ADVANCEMENTS IN HIGH THROUGHPUT PROTEIN PROFILING USING SURFACE ENHANCED LASER DESORPTION/IONIZATION TIME OF FLIGHT MASS SPECTROMETRY

A Thesis
Presented to
The Academic Faculty

by

Vincent A. Emanuele II

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2010

ADVANCEMENTS IN HIGH THROUGHPUT PROTEIN PROFILING USING SURFACE ENHANCED LASER DESORPTION/IONIZATION TIME OF FLIGHT MASS SPECTROMETRY

Approved by:

Xiaoli Ma, Committee Chair
School of Electrical and Computer Engineering
Georgia Institute of Technology

G. Tong Zhou, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Brian Gurbaxani
National Center for Emerging and Zoonotic Infectious Diseases
Centers for Disease Control and Prevention

Arthur Koblasz
School of Electrical and Computer Engineering
Georgia Institute of Technology

Facundo M. Fernandez
School of Chemistry
Georgia Institute of Technology

Date Approved: 12 November 2010

To my parents who raised me to the best of their ability and with an abundance of love.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. G. Tong Zhou and my committee member Dr. Brian M. Gurbaxani for their financial support and guidance during the course of my PhD. Their hard work gave me the opportunity to conduct my research at the Centers for Disease Control and Prevention (CDC) among medical doctors, chemists, microbiologists, molecular biologists, and statisticians: a true inter-disciplinary environment. Without their support, my PhD would not have been possible.

I would like to thank my dissertation committee members: Dr. Xaoli Ma, Dr. Arthur Koblasz, and Dr. Facundo M. Fernandez for taking the time out of their busy schedules to serve on my committee.

There have been numerous other people who have inspired me and influenced me in positive ways over the years.

I would like to express gratitude towards my colleagues at the CDC: Dr. Gitika Panicker, Dr. Beth Unger (current branch chief), Dr. William Reeves (former branch chief), Dr. Toni Whistler, Dr. Sally Lin, and the entire Chronic and Viral Diseases Branch.

I was fortunate to have the opportunity to work for some time with the Computational Systems Biology Lab (CSBL) at the University of Georgia under the direction of Dr. Ying Xu. In addition to Dr. Xu, Dr. Victor Olman of CSBL was undoubtedly a significant influence on my approach to inter-disciplinary research collaborations. I am thankful for discussions with my former lab mates at CSBL: Chindo Hicks, Kyle Ellrott, Juntao Guo, Hongwei Wu, Fenglou Mao, Fengfeng Zhou, and Zhengchang Su.

A debt of gratitude is due to all the current and former members of Dr. Zhou's research group, but especially: Bob Baxley, Ning Chen, Raviv Raich, Lei Ding, Chunming Zhao, Kun Shi, Hua Qian, and Chunpeng Xiao. A very special thanks goes out to Thao Tran who accompanied me into research areas vastly different from the rest of the group and provided me with encouragement, support, and insightful discussions along the way.

Last but not least, none of this would have been possible without the sacrifices and risks taken by my grandparents: Rose, Joaquin, Vincent[1], and Filomena. They left behind everything in their home countries in search of a better way of life and dedicated their whole lives to providing for their families. In turn, my parents put their best efforts into raising my sister and I and supporting our dreams and aspirations.

And to the rest of my family and friends, thank you all for your love and support.

[1]The "original" Vincent on the family tree

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI) is one of several proteomics technologies that can be used in biomarker discovery studies. Such studies often have the goal of finding protein markers that predict early onset of cancers such as cervical cancer. The reproducibility of SELDI has been shown to be an issue in the literature. There are numerous sources of error in a SELDI experiment starting with sample collection from patients to the signal processing steps used to estimate the protein mass and abundance values present in a sample.

This dissertation is concerned with all aspects of signal processing related to SELDI's use in biomarker discovery projects. In chapter 2, we perform a comprehensive study of the most popular preprocessing algorithms available. Next, in chapter 3, we study the basic statistics of SELDI data acquisition. From here, we propose a quadratic variance measurement model for buffer+matrix only spectra. This model leads us to develop a modified Antoniadis-Sapatinas wavelet denoising algorithm that demonstrates superior performance when compared to MassSpecWavelet, one of the leading techniques for preprocessing SELDI data. In chapter 4, we show that the quadratic variance model 1) extends to real pooled cervical mucus QC data from a clinical study, 2) predicts behavior and reproducibility of peak heights, and 3) finds four times as many reproducible peaks as the vendor-supplied preprocessing programs.

The quadratic variance measurement model for SELDI data is fundamental and promises to lead to improved techniques for analyzing the data from clinical studies using this instrument.

CHAPTER I

INTRODUCTION - PROTEOMICS AND BIOMARKER DISCOVERY

1.1 Introduction

Mass spectrometry is one of the more promising tools for solving the fundamental proteomics problem of studying expression levels of proteins in organisms. Proteomics, the study of the complete set of proteins in a living organism, promises to solve many open problems in medicine and public health in the 21^{st} century.

In collaboration with the Centers for Disease Control and Prevention (CDC) in Atlanta, GA, we investigate the capability of high-throughput surface-enhanced laser desorption/ionization (SELDI) time-of-flight (TOF) mass spectrometers (MS) for solving the fundamental proteomics problem. While the SELDI-TOF MS platform does not have the best mass accuracy, sensitivity, and resolution properties compared to other mass spectrometers, its price/performance ratio makes it an attractive tool for high throughput hypothesis generating studies. Additionally, its automated chip chemistry technology enables high throughput analysis of hundreds of spectra per day, a significant virtue.

When receiving raw spectra returns from the SELDI-TOF MS, there are numerous computational steps that need to be performed to process the data to detect the mass/charge (m/z) of the proteins present in the sample and then to extract estimates of their expression levels. These processing steps typically include

- Calibration,
- Noise Filtering (Smoothing),
- Baseline Removal,
- Peak Detection,
- Normalization,

- Peak Alignment.

For each of these processing steps, there are many possible ways to proceed. One of the major difficulties is that there is significant interaction between the processing steps. This means that a failure in one of those steps can corrupt the SELDI-TOF MS signal and bias the interpretation of the entire experiment. Currently, many scientists are struggling to sort through the many choices of software packages that perform these processing steps using a wide array of different approaches. This is a source of confusion and a bottleneck for scientific discovery for investigators at the CDC, who are using SELDI-TOF MS to study the pathogenesis of various diseases.

1.2 The Fundamental Proteomics Problem

The field of proteomics promises to revolutionize the diagnosis and treatment of diseases in the 21^{st} century. One interesting case study is the early diagnosis of cancer. It turns out that most cancer patients are diagnosed in the late stage. For example, [42, 50] point out that 72% of lung, 57% of colorectal, and 34% of breast cancer patients are diagnosed in the late stage. Late-stage diagnosis can often lead to fatality. However, the survival rate is 85% for cancer patients who have been diagnosed in the early-stage. The proteomics approach to this problem is to study the proteins present in easily accessible fluids such as urine, serum, plasma, and mucus to identify *biomarkers* that may be used to detect the onset of the disease in the early stage. The idea is that proteins active in the cells of organs filter into these fluids, especially serum and plasma, enabling a medium for accessible monitoring of organ function.

We now introduce the mathematical formulation of the fundamental proteomics problem. In the fundamental proteomics problem, we have N_C early stage cancer patients and N_H healthy patients (the control group). For each patient, we observe the activity level (called expression level) of P proteins of interest. We can summarize our observations with two matrices. First we have the $P \times N_H$ matrix X_H, with the $(i, j)^{th}$ entry containing the expression level of the i^{th} protein in the j^{th} healthy patient. Similarly, we can group the measurements from the cancer patients into a matrix denoted as X_C. The matrices

X_C, X_H are sometimes referred to as *peak expression matrices* [21], analogous to the microarray data analysis problem. From the clinical data, $\{X_C, X_H\}$, we wish to infer a prediction function $g : \mathbb{R}^P \mapsto \{C, H\}$ such that when we take the corresponding protein expression level measurements of a new patient $\mathbf{x}_\bullet = (x_1, \ldots, x_P)$, we can produce a quality diagnosis $g(\mathbf{x}_\bullet; X_C, X_H)$. The essential properties of g are

- high prediction accuracy and ability to generalize well to the population;
- ability to identify several proteins that are the most important indicators.

While, for clarity of formulation, we have described one of the patient groups as cancer patients, this formulation is by no means restricted to this setting and in fact is very general, encompassing a large number of proteomics problems of interest to clinicians.

Within the framework of the fundamental proteomics problem, a relevant and important scientific problem is that of how to decide which P proteins are interesting and how to measure their corresponding expression levels. There are many possibilities for measuring expression levels, but the two most promising approaches that have been under active development in the last decade are DNA microarrays [105] and mass spectrometry [101] techniques. The SELDI-TOF MS platform is one type of mass spectrometer that is used in the framework of the fundamental proteomics problem. When SELDI-TOF MS or matrix-assisted laser desorption/ionization TOF MS is used to study proteins as stated in the fundamental proteomics problem it is often referred to as protein profiling.

1.3 SELDI-TOF Mass Spectrometry

1.3.1 Motivation for using time-of-flight analysis

Currently there is no standard mass spectrometer that is the universal solution to all possible investigations under the umbrella of the fundamental proteomics problem. There are many different architectures of mass spectrometers, with each design presenting a trade-off in mass accuracy, resolving power, sensitivity, and dynamic range (see [32, 101, 122] for an overview). MS platforms also vary significantly in their cost and throughput abilities. The scientific staff at the CDC is interested in the SELDI-TOF MS platform for the following reasons:

1. Robot-automated sample preparation,

2. High-throughput capability (hundreds of spectra produced per day),

3. Convenient size,

4. Reasonable price/performance ratio.

These virtues give SELDI-TOF MS the potential for future clinical deployment with further advancements. SELDI-TOF mass spectrometers are essentially a variant of matrix-assisted laser desorption and ionization time-of-flight (MALDI-TOF) mass spectrometers, which were pioneered by [66, 107]. Tanaka [107] shared the Nobel Prize for chemistry in 2002 for this ground-breaking work on laser-based soft ionization techniques. The principal difference is that SELDI-TOF mass spectrometers use proprietary chip chemistry to bind specifically to proteins with certain chemical/physical properties to achieve sample complexity reduction [46]. Further, SELDI-TOF MS has a shorter flight tube than most MALDI-TOF MS machines, which results in poorer resolving power. However, this trade-off allows SELDI-TOF MS to be of reasonable geometric size with the potential to be deployed in a clinical setting.

1.3.2 Physical Principles of SELDI-TOF Mass Spectrometry

Here, we present an overview of the sample preparation procedure and physics that enable SELDI-TOF MS to be used for protein profiling. For additional details, see [46]. First, the biological sample of interest is applied to the proprietary ProteinChip array that binds proteins with specific physical/chemical properties. Next, proteins that have not bound or have bound weakly are washed away from the chip, leaving only the desired analytes. The sample is then crystallized within an energy-absorbing matrix, typically sinapinnic acid, as shown in Figure 1[1]. Note that these steps are all fully automated for SELDI-TOF MS when integrated with the Biomek 3000 package (but not necessarily automated with other mass spectrometry platforms). At this stage, the preprocessing of the sample is complete and it is placed into the sample inlet of the SELDI-TOF mass spectrometer.

[1] http://urology.jhu.edu/research/img/proteomics13.jpg

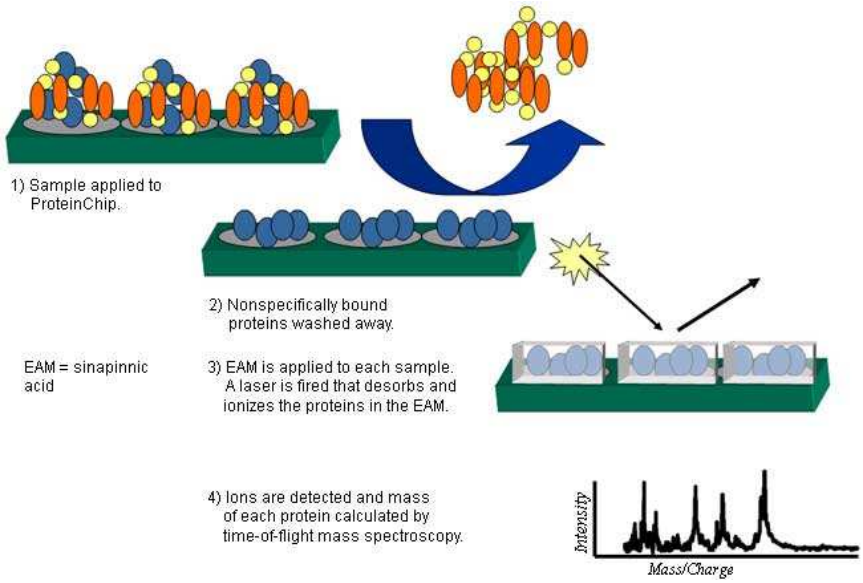

Figure 1: Overview of SELDI-TOF MS sample preparation procedure.

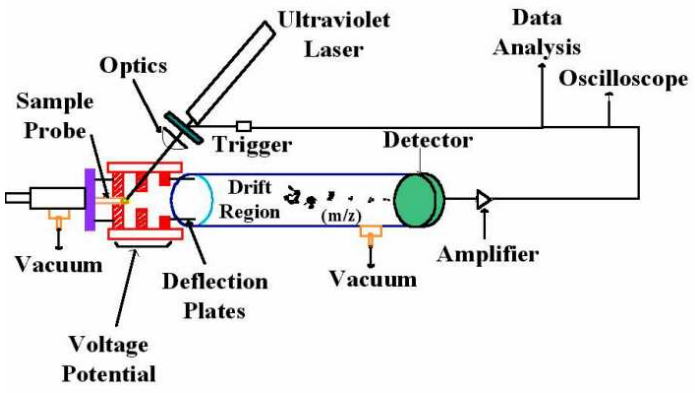

Figure 2: Schematic of MALDI/SELDI mass spectrometry platforms.

A schematic representation of the MALDI-TOF/SELDI-TOF style mass spectrometer is shown in Figure 2[2]. In the MS, a laser is fired at the sample. Upon impact, the matrix absorbs most of the laser's energy, while the protein mixture is both ionized via proton-donation and desorbed from the sample plate into a gaseous state [101]. The remarkable breakthrough in MALDI/SELDI techniques is that most of the protein mixture survives the desorption process intact. Once desorbed, the ions are accelerated through the source extraction region (shown in red in Figure 2) using static electric fields. In principle, all ions have the same kinetic energy as they enter the field-free drift region. Here, the smaller

[2]http://www.psrc.usm.edu/mauritz/images/maldi1b.jpg

ions move quickly toward the detector, while the larger proteins take longer to arrive. For time-lag focusing time-of-flight mass spectrometers, the amount of time it takes to reach the detector is a function of the ion's mass/charge ratio (denoted m/z). The exact relationship is [23]

$$t = \left(\frac{m}{2zE_0s_0}\right)^{1/2} 2s_0 + \left(\frac{m}{2z}\right)^{1/2} \left[\frac{2s_1}{(E_0s_0 + E_1s_1)^{1/2} + (E_0s_0)^{1/2}}\right] + \left(\frac{m}{2z}\right)^{1/2} \left[\frac{d}{(E_0s_0 + E_1s_1)^{1/2}}\right]. \quad (1)$$

In (1), m is the mass of the ion, z is the corresponding charge (often 1 eV), d is the length of the drift region, and s_0 (s_1) are the lengths of the first (second) stages of the extraction region with corresponding electric fields E_0 (E_1). Typically the length of the drift region (d) is much longer than the length of the source extraction region $s_0 + s_1$. Thus, we can simplify (1) as

$$t \approx \left(\frac{m}{z}\right)^{1/2} \left[\frac{1}{\sqrt{2}} \frac{d}{(E_0s_0 + E_1s_1)^{1/2}}\right] \quad (2)$$

$$t \approx a\left(\frac{m}{z}\right)^{1/2} + b \quad (3)$$

Although we theoretically know all the terms in (2) except the m/z, in practice a calibration step is performed by first running a sample containing proteins/peptides of known mass and fitting the constants a, b in (3). For a fixed t, the detector outputs a corresponding intensity value that is roughly proportional to the relative abundance of ions colliding with the detector at time t. Four example SELDI-TOF mass spectra are shown in Figure 3.

1.3.3 Applications of SELDI-TOF MS

The breadth of proteomics-based studies attempted by scientists using the SELDI-TOF MS platform is impressive. Table 1 contains a non-exhaustive list of proteomics applications of SELDI, illustrating the broad reach of mass spectrometry within the framework of the fundamental proteomics problem to solve many problems related to biology, medicine, and public health. The ovarian cancer study by [90] was one of the very early SELDI studies conducted for early detection of cancer. In their paper, the authors predicted a biomarker

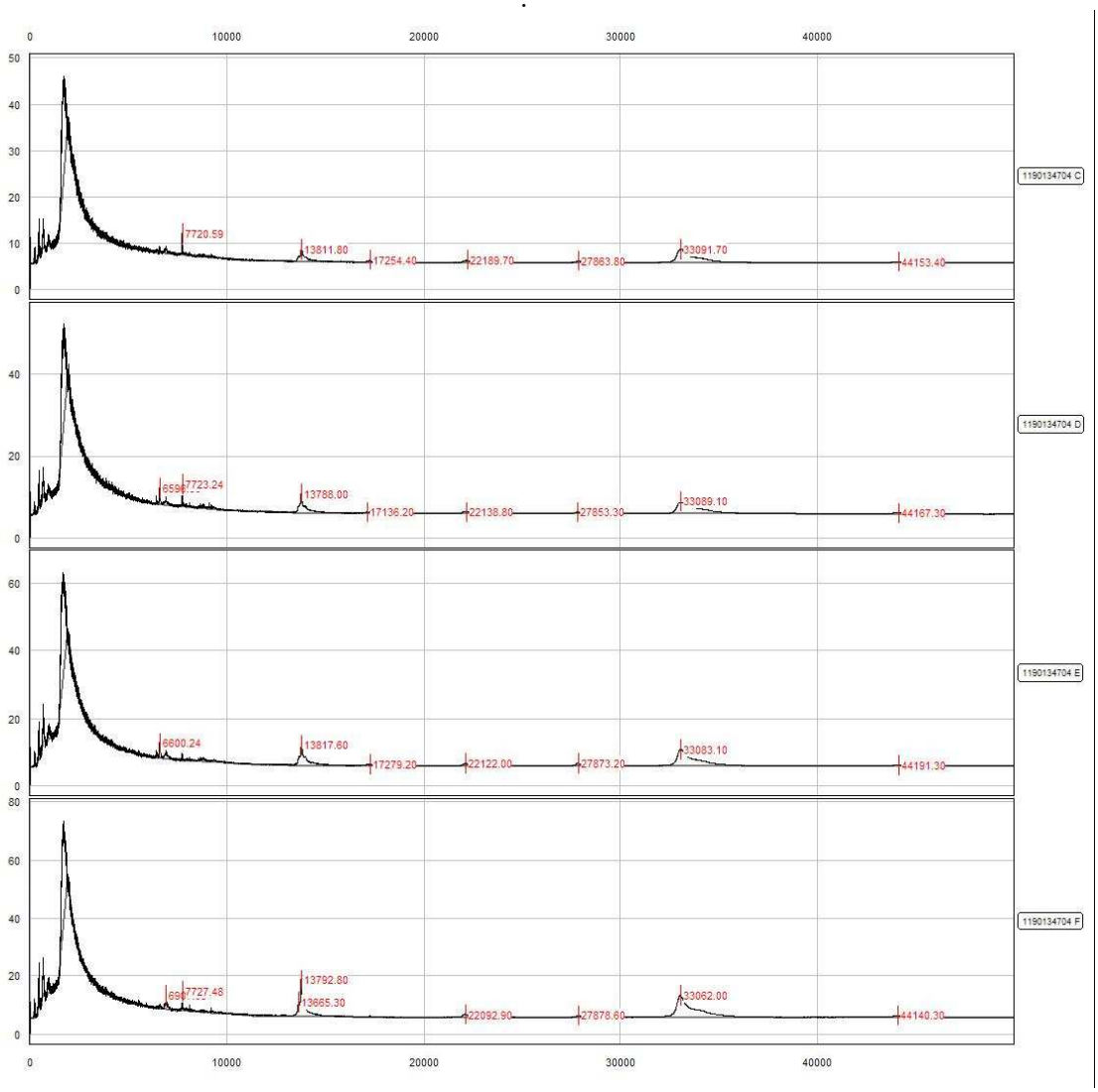

Figure 3: SELDI-TOF MS spectra of blood serum. The x axis is m/z, while the y axis is intensity.

Table 1: Proteomics applications of SELDI-TOF mass spectrometry.

Cancer Studies	
Breast	[20, 67, 89, 93, 134]
Prostate	[1, 9, 84, 94, 100, 115, 132]
Ovarian	[70, 90, 114]
Colorectal	[14, 41]
Nasopharyngeal	[56]
Gastric	[37]
Lung	[129]
Pancreatic	[132]
Subtype Classification	[47]
Other Applications	
SARS	[65, 76, 85]
Species Discrimination	[68, 71]
Rat Liver Cirrhosis	[128, 134]
Schizophrenia	[83]
Bipolar Disorder	[83]
Arsenic and Lead Poisoning	[133]
Rheumatoid Arthritis	[26]
Idiopathic Nephrotic Syndrome	[126]

for early detection that later led to widespread optimism about the capability of the SELDI-TOF MS platform. As will be discussed in Section 1.4, Baggerly and colleagues at the MD Anderson Cancer Center were able to find flaws in the study [5]. Other studies that have received considerable attention in the SELDI community are the prostate cancer study by Adam *et al.* [1] and the breast cancer study by [20].

This brief introduction to SELDI-TOF mass spectrometry is by no means comprehensive. For a deeper discussion of the principles of time-of-flight mass spectrometry, see [23]. For other introductory texts on mass spectrometry, see [52,116]. It is worth mentioning that tandem mass spectrometry (MS/MS), a technique that combines two mass spectrometers in tandem, is an especially active area of research. MS/MS techniques allow researchers to see a more detailed picture of activity, such as post-translational modifications, modulating protein activity in the cell. For a review of some alternative approaches to protein profiling based on MALDI/SELDI, see [32, 55, 104]. Now that we have been introduced to SELDI-TOF MS, we are ready to proceed to the challenging preprocessing steps that must be performed to use this technology effectively.

1.4 Survey of Current Processing Techniques

The key criticism of one of the early major studies performed with SELDI [90] was that an apparent aberration in the preprocessing techniques used by the authors resulted in a significant non-biologically-based bias in the data that explained their ability to predict ovarian cancer patients from healthy patients successfully [5]. The lesson to be learned from this case study is clear: The preprocessing techniques used to remove noise and artifacts from the data are of essential importance and can potentially ruin the conclusions of an investigation. This served as the motivation for developing several alternatives to the Ciphergen ProteinChip/Express preprocessing suite [46] typically purchased with Ciphergen's proprietary SELDI-TOF mass spectrometers. An overview of some of these techniques is presented in this section.

We previously listed the preprocessing steps necessary in Section 1.1. The reason each step is needed will now be made clear, and a brief discussion of existing techniques published for each step will be given.

1.4.1 Calibration

Calibration is needed to account for the practical reality that experiments are run on different days. While care is taken to reproduce all machine settings and sample preparation steps exactly, there is a need to periodically fit machine parameters to a standard sample of known proteins to ensure quality of experimental results.

We have derived a basic form of the calibration equation (3). The Ciphergen vendor-supplied calibration technique adds an additional degree of freedom and considers a calibration equation of the form

$$\frac{m/z}{U} = a(t - t_0)^2 + b \tag{4}$$

where we fit the parameters a, b, t_0 in the usual least-squares way, and U represents a known voltage setting [62]. There are more sophisticated ways of calibrating the SELDI-TOF MS instrument, and we will briefly highlight a few.

Juhasz and colleagues suggest a correction technique after fitting the standard calibration model (3) to the data. Their method includes higher-order terms, $m/z, (m/z)^{3/2}$, with

accompanying parameters estimated by prior knowledge about the distribution of the initial ion velocity off the matrix [64]. Acquiring this knowledge for a particular sample in question requires additional experimental steps. Christian takes this idea and suggests expanding (1) in an infinite series and keeping high-order terms. Then, this higher-order calibration equation is solved using simplex optimization techniques [16].

Rather than drop terms as we did from (1) to (2), Hack and Benner model the residual of the calibrants using a polynomial and use the model of the residual to refine the accuracy of the calibration step [53]. However, one could argue their result to be dubious since they only have five data points to fit in their paper. Bantscheff expands this idea by fitting the residuals (mass errors) of 1800 peptides with a seventh order polynomial, using this to refine the estimates of m/z given by the standard calibration equation [6]. Gobom also uses a similar principle [51]. For the situation when large datasets are being generated, Wolski proposes a way to do calibration without a specified calibration run through the mass spectrometer [123].

Calibration, while very important, is relatively straightforward to carry out. We will see later that the focus of our research efforts should in fact lie elsewhere.

1.4.2 Noise Filtering

SELDI-TOF MS data contains an additive noise component just like many other applications. However, the statistics of the noise processes inherent in SELDI data have not been quantitatively well characterized in the literature. Observing Figure 3, one can propose a reasonable model for raw SELDI data of the form

$$y(t) = b(t) + s(t) + w(t), \tag{5}$$

where $y(t)$ is the observed intensity at time-of-flight t, $b(t)$ is a slow varying baseline effect from the matrix, $s(t)$ is the desired signal (consisting of peaks), and $w(t)$ is an additive noise component accounting for noise in the detector electronics. This sort of model has been suggested by [82]. Noise filtering techniques seek to remove the $w(t)$ component.

One approach to removing noise in SELDI data is through the use of linear time-invariant (LTI) filters. In particular, the use of moving average filters and other low-pass filters has

been proposed [46, 49, 72, 75]. Since mass spectrometry data is spiky in nature (and the high-frequency peaks represent important information), this approach is probably not a good idea.

Other approaches include local regression techniques such as the Savitzky-Golay filter [98, 124]. Bhanot and colleagues use a kernel smoother with a Guassian kernel [9].

Another popular approach for denoising SELDI spectra is using the wavelet transform. In [21, 82], the authors use the undecimated discrete wavelet transform, while [35] applies the continuous wavelet transform on a grid of informative scales and translations.

1.4.3 Baseline Correction

Baseline correction techniques seek to remove the slowly varying baseline $b(t)$, characterized by a sharp increase at low m/z followed by a gradual decay to a constant level, as seen in Figure 3. Removal of the baseline is important for the accurate estimation of protein expression levels for proteins with m/z less than 10,000 Da. Almost every approach first tries to estimate $b(t)$, and then subtract the corresponding estimate from the observed signal $y(t)$. Typically baseline removal is performed after noise filtering.

To estimate $b(t)$, the Ciphergen software [46] uses a convex hull algorithm to estimate the baseline signal as a piecewise linear function. Baggerly and colleagues [4] propose subtracting locally the median or mean of intensities in a window and also allude to the semi-monotonic baseline approach used later in [20, 21]. Sauve and Speed suggest using the top-hat operator from mathematical morphology for baseline removal [97]. Another approach is to fit a loess curve through all the local minima to estimate the baseline, which is carried out by [108]. Andrade and Manolakos [2] use Guassian mixture models for the data in a small window, essentially using the mean of the Guassian that is smaller as the estimate of the baseline. Ressom *et al.* [95] use a spline to regress the baseline in a sliding window.

One particularly unique approach to the problem is given by [72]. The authors propose a charge-accumulation model for the baseline and invert it to subtract the baseline. This is the only study that attempts to model the baseline in terms of the physics of the phenomena,

although the theoretical justification for this model is unclear.

1.4.4 Peak Detection

Finding peaks in SELDI spectra is a critical processing step, as the location of the peak reflects the average mass of the corresponding ion that was detected [101]. At the peak-finding step, we assume that both the baseline and noise components of the SELDI-TOF MS signal have been removed, leaving us with an estimate for $s(t)$, the signature of the proteins in our sample. Peak detection is a general problem encountered in many disciplines. Almost all peak-finding methods contain two parts: finding all local maxima and performing a thresholding test.

Detecting the local maxima is the simplest part of this process. One possible approach is to take a point t to be a peak if $y(t)$ is the largest value in a window $[t-\epsilon, t+\epsilon]$. Such an approach forms the core of [9, 49, 108, 110, 131]. Alternatively, one can take the first difference signal $y'(t) = y(t+1) - y(t)$ (assuming $t = 1, \ldots$ are discrete) and look for zero crossings. Obviously, if one checks the sign of $y'(t)$ before and after the zero crossing, the location of all maxima becomes evident. This strategy is used by [18, 82].

Most of the sophistication in peak finding goes into the calculation of a good threshold test to discriminate a peak that is likely a good signal peak from a spurious peak from the additive noise process. Assuming the noise component is additive white Guassian noise, the best way to do this is to use a matched filter (matched to the shape of the peak) and then apply a threshold to the output [92]. Unfortunately, the statistics of the noise process are non-stationary and non-Guassian, and the shape of the peak is dependent on the resolution of the machine at m/z of interest, the isotopic distribution of the protein at this m/z, and the charge of the protein [101]. All of these factors are unknown to us a priori; thus heuristic threshold tests are used.

We present briefly the peak-detection scheme used in caMassClass [110] to give the flavor of approaches used for SELDI-TOF MS data. The package caMassClass declares a peak at t if

1. $y(t)$ is a local maximum in a window,

2. The moving average filtered version of y evaluated at t is greater than a threshold,

3. A locally calculated z-score is greater than a threshold.

The user must set the size of the window to be used for the moving average filter, the size of the window used for noise statistic estimates, the intensity quantile threshold, and the z-score cutoff. This is a considerable parameter space to search, especially if large numbers of spectra are to be processed.

Some techniques attempt to combine both peak detection and signal estimation/thresholding into one step. For example, in chromatography, Vivo-Truyols and colleagues use the Savitzky-Golay filter in a window and declare the maximum a peak if the second derivative indicates sufficient concavity [112, 113]. Du and Lin perform their peak detection in the wavelet domain and look for "ridge lines," or peaks that occur at the same location across many scales, as an indicator of a peak [35].

There are some references in the IEEE literature that address the peak-detection problem as well, such as [44,63]. However, the ideas on how to proceed are quite similar to those already discussed.

1.4.5 Normalization

Once we have a set of m/z values corresponding to peaks (proteins), we wish to make estimates of the expression levels (intensities). Normalization is an important step in this process, as it removes bias created from variations in the amount of sample put into the SELDI-TOF MS before analysis. Typically we are interested in processing a group of spectra assumed to be generated from the same underlying distribution (e.g., as in the fundamental proteomics problem mentioned in Sec. 1.2). Accounting for several spectra, we need to adjust our model (5) to account for variations in the absolute quantity of proteins that generated each spectrum:

$$y_i(t) = b_i(t) + M_i s(t) + w_i(t), \quad i = 1, \ldots, N_{H \text{ or } C}. \tag{6}$$

In (6), $b_i(t)$ accounts for differences in baseline trends per spectrum, $w_i(t)$ represents different sample paths of the detector noise process (all assumed to be uncorrelated from each

other), and M_i is a scalar factor that accounts for different absolute levels of protein sample delivered to the MS. Note that (6) assumes that relative abundances of the individual proteins in each sample are the same for all i. Normalization techniques seek to remove the effect of M_i.

Suppose we have removed the baseline signal from all of the spectra in (6). This leaves us with

$$\hat{y}_i(t) \approx M_i s(t) + w_i(t). \tag{7}$$

Integrating (7) with respect to all of our observations, we see that

$$\int \hat{y}_i(t)dt \approx cM_i, \tag{8}$$

where we have assumed that the noise processes $w_i(t)$ are zero mean, mean-ergodic random processes [88]. Under these assumptions, it is clear from this development that we can remove the effect of M_i by multiplying the i^{th} baseline removed spectrum (7) by

$$M_i = \frac{\omega}{\int \hat{y}_i(t)dt} \tag{9}$$

where ω is a constant factor we are free to choose. This approach to normalization is called total ion current (TIC) normalization [46]. Dudoit *et al.* [36] investigated the usefulness of TIC normalization and have preliminary results confirming its utility in SELDI data analysis. Other authors [4, 21, 82] also use the TIC method for quantification. Ressom *et al.* [95] also use TIC normalization, additionally choosing ω so that the maximum intensity across all spectra is equal to 100.

Alternatives to TIC have also been proposed. In particular, Tibshirani *et al.* [108] linearly map the 10^{th} and 90^{th} percentiles for each spectra to 0 and 1, respectively. A less robust variant of the Tibshirani approach is used in [9]. Carlson *et al.* [12] proposes a normalization technique similar to an approach used in microarray normalization. First, they focus on several peaks that have large intensity consistently across spectra. Then they use an EM algorithm to estimate the scale factors M_i. They compare their approach to TIC normalization and suggest that the TIC method introduces significant bias. In [130],the authors use the TIC method but first they insist on transforming the intensities by a logarithmic transformation.

1.4.6 Peak Alignment

Peak-alignment algorithms seek to correct for variation in the t (or equivalently m/z) value registered at the detector for the *same* protein. For example, if we were to run 10 pure samples of the protein hemoglobin on 10 consecutive days, we would find that the m/z values detected on those 10 days would be slightly different. This is also evident in Figure 3. In these four example spectra, there is a protein represented with a peak at $33091.70, 33089.10, 33083.10$, and 33062.00 Daltons respectively. While these four m/z values are all different, they represent a protein with the same underlying m/z in this case. There are many factors that go into this, but the most important one is the frequency of calibration runs in the SELDI-TOF MS. Now, imagine we have 100 spectra generated from the same sample, each with $100 - 200$ peak predictions! Most m/z values in this master list will be unique because of the calibration error. Now, since in the fundamental proteomics problem we wish to find P proteins of interest, we need an algorithm to look at the peak predictions on all 100 spectra and produce a list of m/z values that indicate average locations of proteins that are interesting in the sample. The goal of peak alignment is exactly this.

The simplest way to tackle this problem is to interpolate all the spectra onto the same t grid and then perform peak detection on the mean spectrum. This was suggested by [82], and this is also suggested in the documentation for PROcess [49]. Using the mean spectrum in this way has the possibility of eliminating peaks with low expression levels, however. Summarizing briefly, other approaches use hierarchical clustering [108], dynamic programming [97], cross correlation analysis [72, 124, 125], quadratic or cubic splines [62], and interval masking [20, 130, 131].

1.5 The Fundamental Paradox of SELDI-TOF MS Protein Profiling

In Section 1.3.3, we mentioned several early studies [1, 90] that led to widespread optimism about the use of SELDI-TOF MS as the right platform for solving the fundamental proteomics problem. We also mentioned that [5] pointed out inconsistencies in the data in [90]

because of failures in the processing of the samples and the implementation of the data analysis steps described in Section 1.4. This led to increased scrutiny and criticism [13,27–30,85] of studies that were using SELDI-TOF MS. The implications of these studies and accompanying papers criticizing them were clear: it is essential to be especially careful when selecting the parameters for the preprocessing steps described in Section 1.4 to ensure high-quality estimation of the P proteins of interest and their expression levels.

Recalling the formulation from the fundamental proteomics problem described in Section 1.2, we are interested in analyzing N_H spectra from our healthy patients and N_C spectra from the cancer group. If N_H and N_C are very small, then it's possible for an expert chemist/clinician to do the peak calling and peak alignment practically manually. To estimate the expression levels, the clinician could take an iterative approach, trying different parameter settings for baseline removal and smoothing until the result was satisfactory by his/her expert opinion. This is the most straightforward way to produce a high-quality result that avoids the sort of criticism that was alluded to in the beginning of this section.

Typically, for samples generated from blood serum, the number of proteins of interest that clinicians are interested in keeping track of is in the ballpark of $P \geq 50$. In the lingo of statistical pattern recognition, this means that the dimension of our feature space is at least 50. It is a well-known result, called the curse of dimensionality, that as the dimension of the feature space increases, the number of training samples needed to develop a good prediction rule increases exponentially [54]. For example, consider the case when we constrain ourselves to only look at the expression levels of one protein. Now, suppose we consider measuring the expression level of 10 samples of this protein to be a sufficiently dense sampling of expression levels. To get an equivalently good representation of the behavior of 50 proteins, which may be interacting with each other, we would need 10^{50} samples (adapted from an example in [54]). Thus, since the dimension of our feature space is high, we need $N_H \gg P$, $N_C \gg P$ samples. This is just the beginning. Suppose we generate three replicates per patient, with each replicate undergoing a six fold fractionation step, and suppose we are performing a detailed study of cancer with five subtypes and three progressive stages of interest. In this scenario, which is closer to the approach desired by laboratory scientists,

we have compounded the number of classes we are studying by adding fractionation and subtypes and stages of cancer. If we choose to target $P = 50$ proteins of interest for this problem and choose 100 samples for each class to be sufficient for learning the classification rule, then we would be generating $5 \cdot 3 \cdot 6 \cdot 3 \cdot 100 = 27,000$ spectra total that need to be preprocessed to estimate the P proteins of interest and their expression levels!

Currently, no algorithm/software package available has established an ability to preprocess $27,000$ spectra in a fully automated fashion such that it will avoid adding significant bias into the data and repeat the pitfalls of studies such as [90]. To summarize, the scientific community faces a great paradox in the application of SELDI-TOF MS as a candidate solution to the fundamental proteomics problem:

The Fundamental Paradox

- For high-quality, reproducible preprocessing of SELDI-TOF MS spectra we need to keep N_H and N_C small to allow for a large amount of expert intervention in the processing steps

- To estimate an accurate decision rule g discriminating cancer from healthy patients, we need to have N_H and N_C be as large as possible to minimize the effect of the curse of dimensionality and prevent over-fitting the data. This in turn introduces significant bias in the data because of the inconsistent quality of current preprocessing algorithms.

CHAPTER II

BENCHMARKING CURRENTLY AVAILABLE SELDI-TOF MS PREPROCESSING TECHNIQUES

2.1 Abstract

Motivation: SELDI protein profiling experiments can be used as a first step in studying the pathogenesis of various diseases such as cancer. There are a plethora of software packages available for doing the preprocessing of SELDI data, each with many options and written from different signal processing perspectives, offering many researchers choices they may not have the background or desire to make. Moreover, several studies have shown that mistakes in the preprocessing of the data can bias the biological interpretation of the study. For this reason, we conduct the first large scale evaluation of available signal processing techniques to establish which are most effective. We use data generated from a standard, published simulation engine so that "truth" is known.

Conclusions: We select the top algorithms by considering two logical performance metrics, and give our recommendations for research directions that are likely to be most promising. There is considerable opportunity for future contributions improving the signal processing of SELDI spectra.

2.2 Introduction

Mass spectrometry continues to be aggressively pursued as a promising tool for disease biomarker discovery. Currently, there are many possible mass spectrometry platforms one could choose from such as SELDI-TOF, MALDI-TOF, FT-ICR, Ion Traps, Orbitraps, and other popular platforms (reviewed in [32]). Among the many choices available, some biologists have turned to SELDI-TOF MS for two principal reasons:

1. Robot-automated sample preparation using the Biomek® 2000 system,

2. High throughput generation of hundreds of spectra per day.

Early work with SELDI-TOF MS in 2002 produced biomarker predictions and diagnostics for ovarian cancer [90], prostate cancer [1, 94], and led to considerable optimism about the utility of this platform [60, 61, 121, 127].

With increased excitement came increased scrutiny, and in particular, scrutiny of the ground breaking ovarian cancer study by Petricoin and colleagues [90]. After careful examination of the data independently by Sorace and Zhan [103] and Baggerly *et al.* [5], both studies found systematic biases in the data that explained the ability to classify cancer from normal. Some of the key conclusions from [5] indicated problems with baseline correction, inconsistencies in sample preparation techniques used, calibration, feature detection (reproducibility of protein m/z values detected), and noise processes changing in the dataset. This in turn raised alarms throughout the scientific community, leading to numerous articles criticizing SELDI-TOF MS protein profiling [13, 27–30, 85]. While these articles criticizing SELDI may have deterred some researchers, others saw this as a challenge to make SELDI as reproducible and reliable as possible. Several labs addressed reproducibility of SELDI [38,100,120], most notably the Semmes' lab at Eastern Virginia Medical School (EVMS) [100]. Semmes has led a multi-institutional effort to establish protocols to ensure minimum quality standards in the data.

In general there are seven computational steps that need to be performed to extract the desired information from a SELDI-TOF experiment before classification techniques can be

used to identify biomarker candidates. A comprehensive review of all the different possibilities for each preprocessing step is beyond the scope of this paper. For a light overview of why each of the preprocessing steps are needed, see [91]. We show the typical signal processing logic used in Figure 4. Recently, [80] showed that choice of the normalization

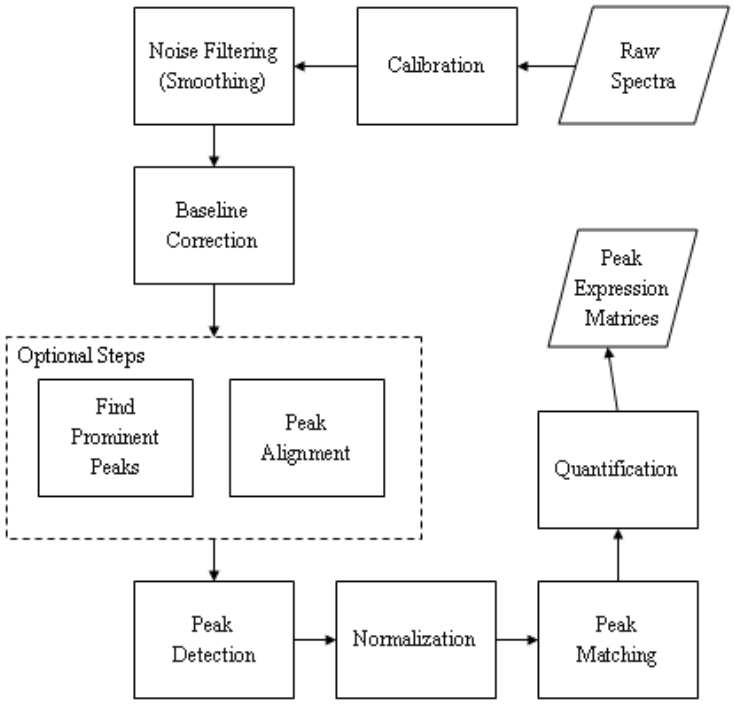

Figure 4: Typical preprocessing procedure for MALDI/SELDI protein profiling data.

step alone can have a significant effect on the quality of SELDI data, from the perspective of both intra-class coefficient of variation and classification accuracy between disease and non-disease classes. From the practicing clinician's point of view, one would want to know which available software package contains the "right" set of preprocessing techniques to ensure the best quality and reproducibility in the data.

Beyer *et al.* [8] have compared two of the major preprocessing software suites. While this was a good first step, the authors only compared two of the possible algorithms. There are several papers proposing new algorithm suites, but in each case the authors choose to compare their new algorithm to only one or two other algorithm, across different datasets (e.g. - [35, 82]).

We propose a significant comparison of the current techniques. In particular, we want to know if any of these approaches show promise as an automated method to preprocess large amounts of spectra reliably. We have carefully chosen among the most popular preprocessing software suites that a practicing scientist would be likely to try out. Approximately half of the programs we have chosen have been used in an applied SELDI study, indicating an immediate need to know their performance capabilities. For example, Ciphergen Express is used in [69] and numerous other studies, PPC in [17], Bioconductor PROcess in [11], and caMassClass in [57]. For a description of our criteria for inclusion in the study, see the supplementary info. The complete list of preprocessing packages we will be examining in this study is shown in Table 2.

Recently, Cruz-Marcelo *et al.* [24] published a study comparing five algorithms, four of which are included in this study [35,46,49,82]. They provided an analysis of peak detection on simulated data, and peak quantification on real data using human control serum. In their work, they found that the vendor supplied software, Ciphergen Express [46], performed well. They also found that [35] and [82] were fairly adept at peak detection. One of their recommendations is that multiple programs be used together to process the data ([35] for peak detection, and [49] for quantification). In this paper, we analyze nine algorithms, listed in Table 2, with the goal to understand in greater detail the processing steps and their effect on the critical task of peak detection. Specifically, we study how peak detection performance may vary with protein mass, protein concentration, and consistency of occurrence in spectra. We find that peak detection performance is highly dependent on mass, concentration, and consistency of appearance. We also find that even the best algorithms leave considerable room for improvement, especially below 10 kDa.

2.3 Materials and Methods

2.3.1 Datasets

Further complicating the comparison of algorithms' performance is that on real SELDI data, we do not actually know what truth is *a priori*. We may detect a peak at 5 kDa, but it is rather difficult to distinguish this peak as a protein of interest, a contaminant, an artifact

Table 2: General information regarding available software for SELDI data processing

Program Name	**Reference**	**Affiliation**	**Year**	**Implementation**	**Platform**
Ciphergen Express	[46]	Ciphergen	2002	Binary	Windows
Cromwell	[21]	MD Anderson, Univ. of Texas	2005	Matlab	Mac/Linux/Windows
Mean Spectrum	[82]	MD Anderson, Univ. of Texas	2005	Matlab	Mac/Linux/Windows
PPC	[108]	Stanford	2004	Excel/R	Mac/Linux/Windows
Bioconductor PROcess	[49]	Multiple	2004	R	Mac/Linux/Windows
PROcess/Mean Spectrum Option	[49]	Multiple	2004	R	Mac/Linux/Windows
GenePattern	[75]	Broad Institute, MIT	2005	R	Mac/Linux/Windows
Bioconductor CaMassClass	[110]	NCICB/NIH	2006	R	Mac/Linux/Windows
MassSpecWavelet	[35]	Northwestern	2006	R	Mac/Linux/Windows

introduced by the preprocessing steps, or a spurious peak due to the additive noise process inherent in SELDI spectra. Because of this, we cannot say that algorithm A is better than algorithm B just because A found a peak in some real data that B did not find.

Therefore, if we wish to make valid scientific conclusions about which preprocessing algorithms show the most promise, we must compare them on data that we know the true protein content of in advance of the experiments. One possibility is to use the calibration samples (or spike-in data), that typically contain a small number of peptides (< 10) of known m/z value. This approach was taken in [35, 106]. However, this approach does not accurately reflect the complexity of the samples typically profiled with SELDI, which often contain on the order of 100's of proteins.

With this in mind, we conclude that the only reasonable way to compare the preprocessing algorithms is by using a model of the dual stage delayed-extraction TOF architecture typical of SELDI platforms. Fortunately, Coombes and colleagues have developed a low resolution MALDI/SELDI MS simulation engine from first principals [18, 82]. The principal virtues of their simulation engine are:

1. It is based on the physical principles of the dual stage delayed-extraction MALDI architecture

2. Simulation parameters are estimated from actual low resolution MALDI experiments

3. It allows for generating data from complex mixtures of virtual protein populations, where we know the true protein m/z value at all times.

We use 100 different datasets representing samplings from the same population, with each sampling containing 100 spectra to measure the algorithms against each other. The data are available for download at `http://bioinformatics.mdanderson.org/Supplements/Datasets/Simulations/index.html`. For detailed discussion of the simulation dataset, see the supplementary information provided or [18, 82].

Originally, [18, 82] proposed their simulated dataset be used as a benchmark for preprocessing algorithms. In lieu of the discussion presented here, we agree and proceed with their dataset for our comparisons.

2.3.2 Performance Comparison

The goal for each algorithm in this study is to reconstruct the list of 150 protein m/z values for each of the 100 datasets (sample populations). In order to assess performance capabilities, we have run each of the algorithms over a wide range of parameter choices as recommended by the developers of each package. For more information about how the parameters were selected, see the supplementary information. For each parameter choice on each dataset, we calculate the observed false discovery rate (FDR) and true positive rate (TPR, also called sensitivity), as follows

$$FDR = \frac{FP}{FP + TP} \qquad (10)$$
$$TPR = \frac{TP}{TP + FN}.$$

Similar to [82], we define the TP (the number of true positives) as the number of the 150 virtual protein m/z values having at least one predicted m/z value within 0.3% relative error. The FP is defined as the number of predicted m/z values not within 0.3% of any of the 150 virtual protein m/z values for this dataset. Similarly, FN is the number of the 150 virtual protein values without any predicted m/z value within 0.3% relative error. As done in [82], we also keep track of predictions and proteins that match a multiple number of times, represented by the quantities $MM1$ and $MM2$.

2.3.2.1 Operating Characteristics

One way to view the peak detection problem is as a special type of binary classification problem [82]. From this perspective, a decision is made at each m/z value whether there exists a virtual protein or not. Because of the huge number of hypotheses being tested, this manifests itself as a multiple testing problem. In this scenario, the classic notion of restricting type I error (e.g. - using Neyman-Pearson tests) is not quite as useful [7]. However, one certainly is concerned with the false-discovery rate, as defined in Eq. 10.

One of the best ways to understand the performance of a program is to look at the trade-offs between sensitivity and false-discovery rate obtained by varying free parameters.

From these empirically observed points, we use loess smoothing to estimate operating characteristics (OC) for each program as a summary of its performance on each dataset. This is similar to the receiver-operating characteristic [43], except the false-discovery rate has been substituted for the type I error rate in our case. Other authors have reported similar operating characteristics in their work as well [35, 82, 106]. This multiple testing problem, along with the use of operating characteristics to address it have been used in the microarray analysis literature also [15].

We wrote numerous scripts to automatically evaluate each program over a wide range of parameter combinations. The simulations benchmarking the algorithms consumed approximately a year of computing time, spread across several cores in a small computing cluster. For the purposes of evaluating the performance of preprocessing programs for SELDI, we have developed our own software toolbox, written in Matlab, and provided as supplementary information. Unfortunately, there is no built-in scripting capability for the Ciphergen Express program. Thus we had to perform analysis with Ciphergen by hand, which was rather laborious. As a result, the Ciphergen Express program contains less operating points and we used a piecewise linear representation of its operating characteristic rather than a loess smoothed one.

2.3.2.2 Metrics for Ranking Algorithms

We propose two principal figures of merit for the purpose of ranking the algorithms from most to least promising. While we are most interested in average performance across the 100 datasets, we do report the standard error as well. Indeed, it would not be a good result if performance varied significantly for different samplings from the same population, as in our simulation set up. We define each metric first, with explanations to follow.

1. MEANTPR: Let m_i be the mean TPR reported for dataset i. Then, $MEANTPR = \frac{1}{100}\sum_i^{100} m_i$.

2. PAUC (Partial Area Under the Curve): For each dataset, calculate the area under the OC curve between FDR values of 0% and 50%. PAUC is the average of this quantity across all 100 datasets.

The MEANTPR metric represents the sort of sensitivity that a practicing biologist is likely to observe if they proceed by trying a few parameter choices in the range suggested by the algorithm's author and look at the result. In other words, MEANTPR reflects the average percentage of real proteins that one would expect to find with the corresponding method being evaluated.

In contrast to MEANTPR, the PAUC metric considers the well known trade-offs between FDR and TPR that are typical when one varies parameters in a hypothesis testing setting. Not all operating points on the OC curve are useful for applications of interest such as biomarker discovery. For example, when a program is operating at a FDR of 50%, it corresponds to the "coin flip" scenario: half of the predicted peak m/z values correspond to true proteins, and half are erroneous, so that operating points with $FDR \in (50\%, 100\%]$ are not very useful. Because of this, we define our second benchmarking metric, the partial area under the curve (PAUC), to be the the area under the OC curve over the domain $FDR \in [0\%, 50\%)$.

2.4 Results and Discussion

2.4.1 Global Ranking of the Algorithms

Using the metrics defined in Section 2.3.2.2, we have ranked the algorithms from most promising to least, as illustrated in Tables 3 and 4. Because the MEANTPR metric relates

Table 3: Algorithm ranks using mean sensitivity (MEANTPR) as the figure of merit.

Rank	Program Name	Metric Score	SEM
1	MassSpecWavelet [35]	51.5%	3.6%
2	PPC [108]	47.2%	3.0%
3	Ciphergen [46]	47.1%	5.3%
4	Mean Spectrum [82]	40.5%	3.1%
5	Cromwell [21]	28.7%	5.4%
6	GenePattern [75]	19.4%	2.1%
7	PROcess [49]	19.0%	1.9%
8	CaMassClass [110]	17.2%	2.1%
9	PROcess/Mean Spectrum [49]	6.1%	1.3%

closely to the way the data analysis usually proceeds in the lab, we place heavy emphasis on the importance of the results in Table 3. Two algorithms, MassSpecWavelet [35] and

Table 4: Algorithm ranks using partial area under the curve (PAUC) as the figure of merit. Note, PPC was excluded since it had no observed operating points for $FDR \in [0, 0.5)$.

Rank	Program Name	Metric Score	SEM
1	Ciphergen [46]	0.284	0.07
2	Mean Spectrum [82]	0.280	0.05
3	CaMassClass [110]	0.279	0.03
4	MassSpecWavelet [35]	0.251	0.05
5	Cromwell [21]	0.206	0.05
6	GenePattern [75]	0.164	0.06
7	PROcess [49]	0.108	0.02
8	PROcess/Mean Spectrum [49]	0.086	0.09

PPC [108], stand out with mean sensitivities of 51.5% and 47.2% respectively.

With respect to PAUC measure, Ciphergen Express [46] and Mean Spectrum [82], are the top two performers. Their corresponding operating characteristics, along with MassSpecWavelet's [35], are displayed in Figure 5. Operating characteristics for the rest of the programs are given as supplementary information.

2.4.2 Potential for Identifying Special Classes of Proteins

While the results in the previous section present us with what seems to be the most promising algorithms, the practical question is whether these programs compliment each other or are redundant. Most importantly, we wish to know what are the characteristics of the virtual proteins found by each algorithm?

2.4.2.1 Dependence on Mass

We have assessed the mean sensitivity of the algorithms as a function of mass, in order to identify regions of the m/z axis where algorithms tend to perform well or poorly. The results of our analysis are summarized in Figure 6. Generally, the trend for all algorithms is that MEANTPR increases with protein m/z. All algorithms experienced difficulties finding the true protein m/z values for proteins less than about 6 kDa. This is rather unfortunate news for the application of protein profiling to many disease investigations, as there may be important small proteins that one may want to find here such as defensins (peptide antimicrobials) in the 3 to 3.5 kDa range.

We have analyzed the effect of peak density (measured in peaks per Da) on the MEANTPR

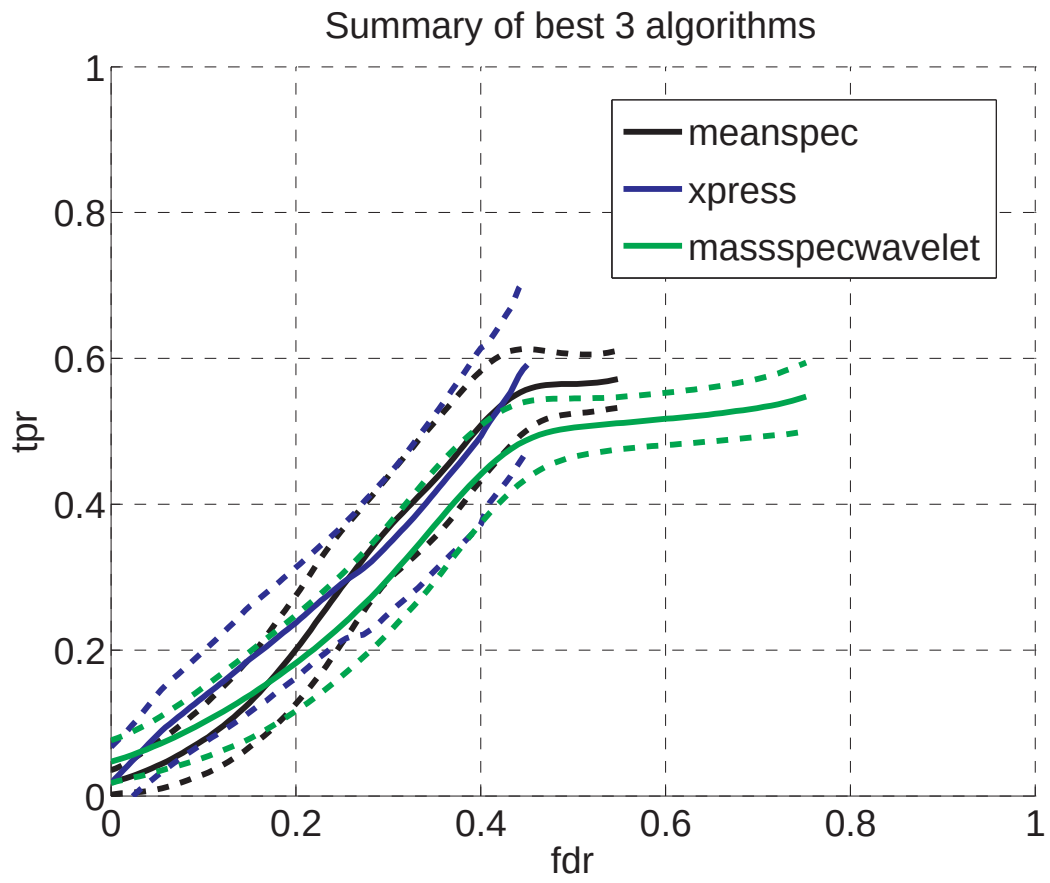

Figure 5: Operating characteristics for the top 3 programs (with respect to PAUC metric). The solid line represents the mean operating characteristic, with the dashed lines indicating the mean plus/minus the standard error.

observed for the algorithms. The results of this analysis are shown in Figure 7. Figure 7 (top) shows that in our simulation model the peaks tend to be more spread out at higher at higher m/z values.

We have observed this to be true in typical blood serum spectra as well. We have further noted a strong negative correlation between peak density and MEANTPR for three of the top performing algorithms. Thus, in our opinion peak density is the true underlying cause for the trends observed in Figure 6.

One may also wonder whether poor calibration may be affecting the performance of the algorithms. The calibrant m/z values used in the study are shown as red dots in Figure 6 along the bottom. Clearly, the calibrants span the m/z range of nearly all protein m/z

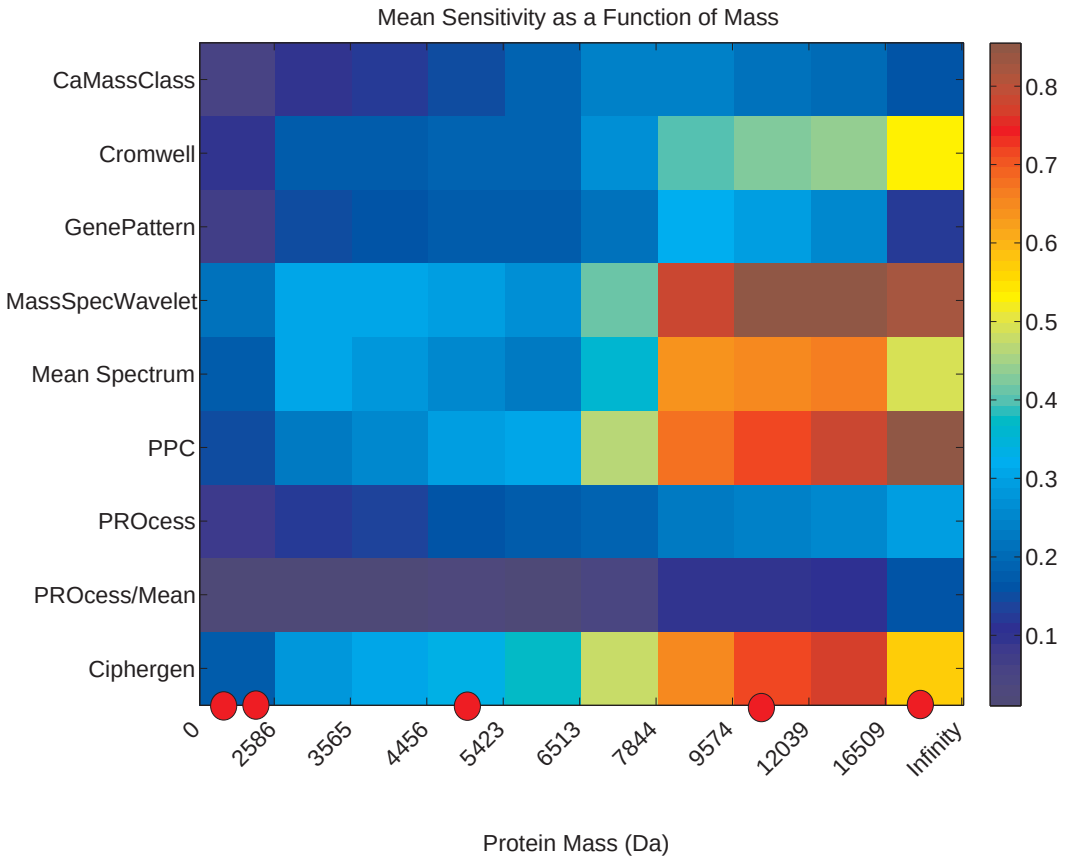

Figure 6: Mean sensitivity as a function of mass for each algorithm. Each mass bin was specially selected in increments of 10% of the quantile function of all the pooled true protein masses from all 100 datasets. The red dots on the x axis indicate the approximate m/z values of the virtual protein standards used to fit the calibration equation (occurring at 1, 2, 5, 10, and 20 kDa).

values used in the simulation, thus we do not believe calibration contributes significantly to our observations in Figure 6. However, calibration does certainly contribute to mass error. In particular, relative mass errors of 0.34%, 0.17%, 0.07%, 0.05%, and 0.01% are observed in single charged virtual proteins with molecular weights of 1, 2, 5, 10, and 20 kDa respectively [19].

There is also evidence that the baseline contributes to biased estimates of the true protein masses at very low m/z values. Figure 7 (top), shows that for the lowest m/z bin, the peaks are not very dense. However, the corresponding MEANTPR for this point is 19% (average over the three algorithms). This confirms our experience with SELDI. Namely,

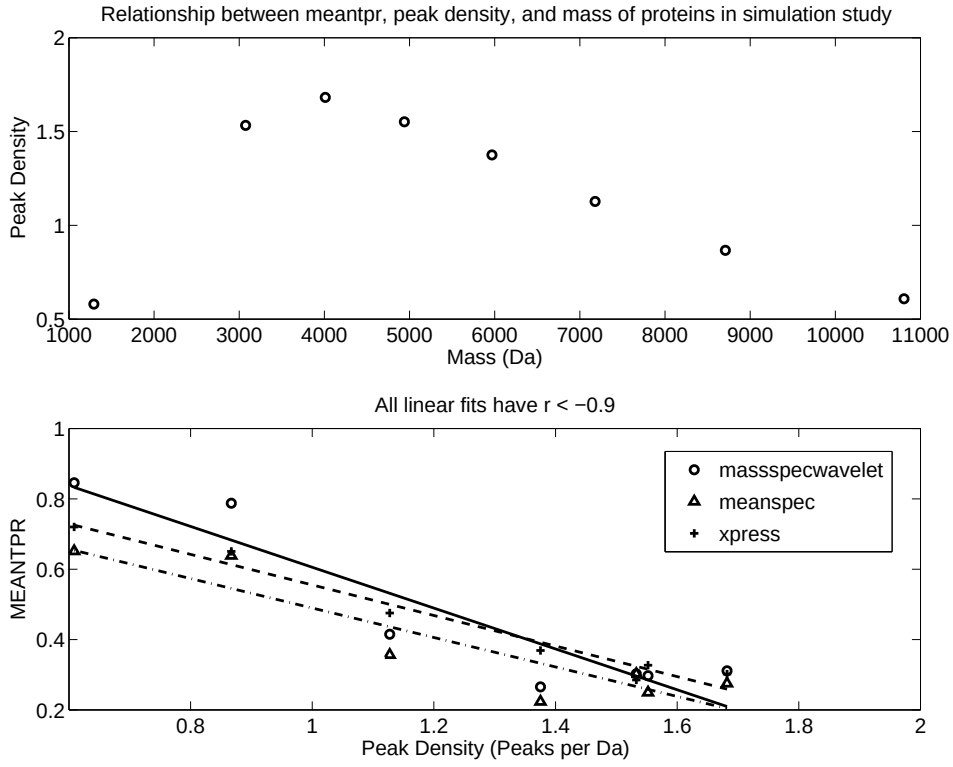

Figure 7: (top) Exploring density of peaks corresponding to proteins (measured in peaks per Da) as a function of molecular mass in the simulation. Clearly, the peaks are more densely packed at lower m/z values. The scale of the x axis is given in units of 10^4 Da. (bottom) Density of proteins is a predictor of algorithm performance. The most dense areas are at low mass, which partially explains why the algorithms performed so poorly in these areas. Further, the abundant proteins also tend to occur at low Da, explaining why we have observed decreased performance for increased abundance simulation parameter.

that there are few peaks appearing around 1100 Da, and that extracting good estimates of the corresponding m/z values in this area are difficult due primarily to interference from the baseline. In Figure 7 (bottom), we have removed this outlier for our corresponding analysis due to the interference effect from the baseline and calibration at m/z around 1100 Da. On the other hand, Figure 6 shows that, for proteins with a mass greater than about 7800 Da, MassSpecWavelet, Ciphergen Express, Mean Spectrum, and PPC are quite effective at detecting them. This is good news, as there are numerous interesting proteins in this size range.

We also uncovered an error in the way the isotope distribution is calculated for each protein m/z value in the simulation engine, which results in a predictable bias in peak

locations. After correspondence with Kevin Coombes of MD Anderson (one of the principal authors of the simulation), the problem was promptly fixed in the current version of the simulation engine available from their website. Since the additional mass error introduced from this problem is typically small (on the order of 0.01 percent of mass) and presumably does not favor any algorithm, we proceeded with the study [19].

2.4.2.2 Dependence on Prevalence

In the simulation model for SELDI data, protein prevalence is the probability that a virtual protein appears in a spectrum. This models inherent randomness observed in SELDI data, as it tends to be rare for a peak corresponding to a single protein to be observed in 100% of the spectra in one's data. We have investigated the effect of prevalence on the performance of the algorithms, shown in Table 5. As expected, the mean sensitivity increases as protein prevalence increases. This confirms intuition, as there are more chances of finding the occurrence of this protein with higher prevalence. MassSpecWavelet [35] uniformly ranks best for proteins of all classes of prevalence, and performs most uniformly with variations in prevalence. This is potentially important, as low prevalence proteins may indicate the presence of in important subgroup of samples within the dataset.

2.4.2.3 Prevalence and Abundance Interactions

Protein abundance is measured as the mean log intensity of the peaks corresponding to a virtual protein in our model. This is related to the modeled protein concentration in the virtual samples. It is estimated that protein concentrations in human cells easily span seven to eight orders of magnitude, with speculation as high as twelve orders of magnitude [22]. Therefore, it is clearly of interest to understand how abundance and prevalence interact and affect the performance of the algorithms investigated. Morris *et al.* conducted such an investigation in [82]. However, they only looked at the performance of two algorithms, namely [21, 82].

We list the top two algorithms for each subclass of virtual proteins, along with their mean sensitivities in Table 6. The table is organized so that each of the 9 cells contains results for roughly the same number of peaks. Again, the performance of the MassSpecWavelet [35]

algorithm is superior in all but the high prevalence/high abundance and medium prevalence/high abundance cases, which favor Ciphergen Express [46]. PPC [108] and Mean Spectrum [82] also do well.

One counter-intuitive observation apparent in Table 6 is that for a fixed prevalence range, mean sensitivity decreases with increasing abundance. The explanation for this is as follows, and is addressed primarily in Figure 7. The simulation parameters, estimated from real SELDI data, specify a negative correlation between abundance (mean log intensity of peak height) and the log of mass of the protein in the simulation. In other words, high abundance proteins occur at the lower end of m/z values in the simulations, where peak density is the dominating factor that complicates peak finding, as discussed earlier. One of the purported benefits of SELDI is its ability to observe hundreds of proteins simultaneously in a complex medium such as blood serum. Our results indicate that this may not be so advantageous, since more crowded peaks are difficult to resolve, even with the best programs (Figure 7 bottom). The fact that peak density can be such a significant factor in peak finding is motivation for higher resolution instruments and improved sample complexity reduction methods.

2.5 Concluding Remarks

We now consider the entirety of our results and analysis. While the performance of all algorithms is generally poorer than was expected, with roughly half of the 150 protein m/z values being successfully recovered by the best algorithms, there is considerable room for optimism and improvement.

Both by quantitative (e.g. MEANTPR as a function of mass, prevalence, and abundance) and qualitative (e.g. usability and intuitiveness) measures, Ciphergen Express [46], MassSpecWavelet [35] and Mean Spectrum [82] have a performance edge over the other approaches. That is, they identify a peak within 0.3 percent of it's true mass in a given spectrum more often and more parsimoniously (i.e. in a simpler way) than the other algorithms. The fact that Ciphergen Express performed well in our analysis should be welcome news to the SELDI community, as this is the most commonly used program by researchers

processing their data.

The best program we tested in terms of user friendliness was Ciphergen Express [46]. While it's performance is among the top programs used in this study, we still recommend analyzing one's data with MassSpecWavelet [35] and Mean Spectrum [82] as well. Therefore, we strongly recommend that biologists make the effort to explore their data using these programs. We encourage biologists with little or no programming experience to collaborate with statisticians, engineers, and computer scientists, as we have found such collaborations to provide a true synergy. An excellent overview of the sources of variation in SELDI data and what can be done to mitigate these effects is given in [96, 119].

There is considerable opportunity for contributions in the area of signal processing for MALDI/SELDI protein profiling experiments from statisticians, physicists, mathematicians, engineers, and computer scientists.

We strongly believe that computational scientists have the same responsibility as laboratory scientists to ensure their results are reproducible by the broader community. We hold the ideas for reproducible computational research suggested by the Claerbout lab at Stanford University as an ideal that the bioinformatics/proteomics community should strive for [99] [1]. Several of the software packages used in this study did not work when they were first downloaded; all of them required various amounts of tweaking to get them going. These hurdles could block a laboratory from using a technically superior package and possibly making a biomarker discovery. Therefore we strongly recommend that developers test their software on other platforms before making them available to the community, and that they support the software once it is made available.

Furthermore, we strongly encourage authors to make code available for download that can easily be run to regenerate as many of the major results presented as figures and tables in their paper as is feasibly possible. While this requires a small amount of extra effort, it serves the community as a whole by speeding up the rate at which discoveries can be made and confirmed. Towards this end, we have made available as supplementary information approximately two gigabytes of Matlab code, perl scripts, and data that can be used to

[1] `http://sepwww.stanford.edu/research/redoc/`

generate the figures and tables used in this publication by running a few simple commands. This is available via FTP from the authors by request.

There are several areas that stand out as needing immediate progress. First, we must find ways to improve performance of the signal processing steps at low m/z values. There are many important proteins with mass under 7 kDa that scientists will not want to miss.

Most signal processing suites explored in this paper took a top-down approach to design. We believe that top-down approaches in this area have reached their potential. For progress to continue in this area, researchers must come up with good models of the data derived from a minimal number of sound assumptions. Malyarenko *et al.* has made some initial progress in this way by using a charge accumulation model for the baseline drift observed in MALDI/SELDI signals [72]. We recommend that more algorithms in the future make use of physical models of the data first.

Looking farther down the road, high throughput, low resolution MALDI/SELDI protein profiling will certainly benefit from a higher level of automation of the signal processing steps. The trend in public health related studies is for the amount of patients/data generated to be ever increasing. With protein profiling studies feasibly having thousands of patients in case and control groups, it will become impractical to manually tweak parameters to ensure each spectrum gets processed just right.

Table 5: Algorithm rankings as a function of protein prevalence. Performance is measured using mean sensitivity (MEANTPR). The 95% confidence interval for the average sensitivity is given in the parentheses.

Rank	Prevalence (p)					
	$0.00 - 0.28$		$0.28 - 0.80$		$0.80 - 1.0$	
1	MassSpecWavelet	48.4% (35.1, 61.7)	(Same)	52.9% (39.0, 66.9)	(Same)	53.5% (39.6, 67.3)
2	PPC	44.1% (33.9, 54.4)	Ciphergen	50.2% (36.7, 63.7)	(Same)	52.9% (38.7, 67.0)
3	Ciphergen	39.6% (33.1, 45.0)	PPC	49.6% (39.4, 59.9)	Mean Spectrum	48.5% (37.6, 59.5)
4	Mean Spectrum	31.4% (22.6, 40.3)	(Same)	42.9% (31.1, 54.6)	PPC	48.0% (38.1, 57.8)
5	Cromwell	26.4% (14.9, 38.0)	(Same)	30.2% (17.6, 42.8)	(Same)	29.5% (17.1, 41.9)
6	PROcess	17.6% (13.1, 22.1)	GenePattern	20.3% (13.1, 27.6)	GenePattern	25.3% (16.9, 33.7)
7	GenePattern	13.5% (7.8, 19.1)	PROcess	20.2% (15.0, 25.5)	CaMassClass	21.8% (15.0, 28.6)
8	CaMassClass	12.0% (7.2, 16.8)	(Same)	18.6% (12.5, 24.6)	PROcess	19.4% (12.8, 26.0)
9	PROcess/Mean	4.8% (2.2, 7.4)	(Same)	6.3% (2.8, 9.9)	(Same)	7.5% (3.7, 11.2)

Table 6: Top two performers for different combinations of prevalence and abundance. Performance is measured using mean sensitivity (MEANTPR).

Prevalence	Abundance (a, mean log intensity)		
p	< 9.07	$9.07 - 9.72$	> 9.72
$0.00 - 0.28$	MassSpecWavelet 51.8%	MassSpecWavelet 48.2%	MassSpecWavelet 45.0%
	PPC 47.7%	PPC 43.1%	PPC 41.3
$0.28 - 0.80$	MassSpecWavelet 57.3%	MassSpecWavelet 52.5%	Ciphergen 50.6%
	PPC 53.3%	Ciphergen 50.0%	MassSpecWavelet 48.2%
$0.80 - 1.0$	MassSpecWavelet 58.7%	MassSpecWavelet 53.4%	Ciphergen 52.2%
	Ciphergen 53.5%	Ciphergen 53.0%	Mean Spectrum 49.0%

CHAPTER III

QUADRATIC VARIANCE MODELS FOR ADAPTIVELY PREPROCESSING SELDI-TOF MASS SPECTROMETRY DATA

3.1 Abstract

Background: Surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI) is a proteomics tool for biomarker discovery and other high throughput applications. Previous studies have identified various areas for improvement in preprocessing algorithms used for protein peak detection. Bottom-up approaches to preprocessing that emphasize modeling SELDI data acquisition are promising avenues of research to find the needed improvements in reproducibility.

Results: We studied the properties of the SELDI detector intensity response to matrix only runs. The intensity fluctuations and noise observed can be characterized by a natural exponential family with quadratic variance function (NEF-QVF) class of distributions. These include as special cases many common distributions arising in practice (e.g.- normal, Poisson). Taking this model into account, we present a modified Antoniadis-Sapatinas wavelet denoising algorithm as the core of our preprocessing program, implemented in MATLAB. The proposed preprocessing approach shows superior peak detection sensitivity compared to MassSpecWavelet for false discovery rate (FDR) values less than 25%.

Conclusions: The NEF-QVF detector model requires that certain parameters be measured from matrix only spectra, leaving implications for new experiment design at the trade-off of slightly increased cost. These additional measurements allow our preprocessing program to adapt to changing noise characteristics arising from intralaboratory and across-laboratory factors. With further development, this approach may lead to improved peak prediction reproducibility and nearly automated, high throughput preprocessing of SELDI data.

3.2 Background

Mass spectrometry is a promising technology for biomarker discovery [59]. There are a wide variety of mass spectrometers from which one could choose from during the design of a biomarker discovery experiment, reviewed in [32]. Matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS, or just MALDI) can ionize whole proteins intact over a wide range of protein mass values, making it suitable for biomarker discovery in complex media such as blood serum, where both protein concentrations and masses vary greatly [22]. Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS, or just SELDI) [58] is a variant of MALDI that adds an on-chip chromatographic separation step at the front end of the analysis pipeline. This, combined with robot-automated sample preparation, enables SELDI to be high-throughput, an attractive feature for many laboratories. For a recent review of the application of SELDI in the context of biomarker discovery, see [10].

The typical SELDI work flow involves the collection of samples (e.g.- blood serum) from patients, application of the samples to SELDI ProteinChips® selected for desired physicochemical properties, and analysis in the SELDI mass spectrometer. The raw data must be preprocessed to detect relevant peaks which correspond to proteins in the sample. Typical signal preprocessing steps performed are spectral alignment, denoising/smoothing, peak detection, peak matching, normalization, and quantification (see Figure 1 of [39]). The preprocessing of the raw SELDI spectra is typically accomplished using one of several available software packages (reviewed in [24, 39, 117]). Artifacts due to insufficient preprocessing of the data have, in the worst case, led to erroneous biological conclusions in early SELDI studies [5, 28, 103]. This fact inspired several important comparison studies of SELDI preprocessing algorithms [24, 39, 80, 117]. We now briefly summarize a few of the major contributions. For a more detailed overview, see the introduction of [39].

Coombes *et al* introduced the use of wavelets for denoising SELDI spectra [21], providing a more adaptive approach to denoise compared to moving average filters (e.g., as in [46]). Meanwhile, Morris *et al* introduced the notion of a mean spectrum, which represents average protein activity of a group of spectra. Under non-restrictive assumptions,

the mean spectrum has less noise and allows one to circumvent complicated peak matching algorithms that consolidate peak predictions among individual spectra into a consensus prediction. Malyarenko *et al* introduced a novel baseline removal algorithm based on a proposed charge accumulation model of the saturation phenomenon of the detector [72]. This was one of the first algorithms that was designed from the "bottom-up", starting with physical considerations of SELDI. Later, deconvolution filters were shown to be a possible approach for improving mass resolution of SELDI [48, 73, 74].

Sköld *et al* analyzed single-shot spectra [102], the basic components of a final SELDI spectrum obtained by summing the results of many laser shots. They suggested that the observed counts in the single shot spectra may be proportional to a Poisson random variable, proposing a heteroscedastic model for the data. Meuleman *et al* also make use of single-shot spectra (sub-spectra) to derive a preprocessing algorithm based on analyzing these components separately [79].

In an attempt to improve on the bottom-up approach to preprocessing, we analyze the statistics of the SELDI signal over a wide range of intensity values. Based on data presented herein, we propose a natural exponential family model with quadratic variance function for the statistics of the detector response for SELDI experiments. We believe this model is a plausible explanation for acquisition of single-shot spectra, summing of single-shot spectra into a final spectrum, and extracting protein estimates from a mean spectrum under a unified framework. Under this framework, we introduce a new preprocessing approach, adaptive to changing noise characteristics per spectrum and per experiment, and show favorable peak prediction performance.

3.3 Results

3.3.1 Buffer-only intensity measurements

Electronic measurements exhibit natural random fluctuations [111]. In many cases, these fluctuations are independent of the signal and are modeled as additive white Gaussian noise. In order to understand the nature of the noise fluctuations inherent to SELDI, we study the response of the detector under controlled experiments applying different buffers instead of

protein samples under varying laser intensities (as in [96]).This eliminates the complexity introduced by adding serum to the chips while facilitating measurements of ion counts over a wide range of intensity values. In principle, this gives us a set of n repeated experiments from which we can study the statistics of the detector response compounded with noise and interference inherent to SELDI. In this fashion, we have generated two separate buffer + matrix datasets, denoted BUFFER1 and BUFFER2, which represent data generated on the *same* SELDI PBS IIc machine by *different* scientists and different machine parameters. BUFFER1/BUFFER2 contain 183/114 spectra, respectively.

We visualize all of the spectra in BUFFER1 and BUFFER2 in Figure 8. In particular, we are interested in analyzing the region between 3 and 30 kDa, since this is the mass focusing region in our experiments. In this region, the observations across spectra for a fixed time (mass) point represent approximately independent, identically-distributed measurements in BUFFER1 or BUFFER2, respectively. Figure 8 shows the median, 75% quantile, and 25% quantile of BUFFER1 and BUFFER2. The median spectrum shows the form of an ordinary measurement, with any measurement between the 75% and 25% spectrum lines considered typical as well.

Figure 8 shows us the behavior of the typical buffer + baseline signal component seen in all SELDI raw spectra. Indeed, we see that changing different machine settings leads to different response properties. For BUFFER2, the median spectral response is large in the range shown, and the distribution of responses is symmetric about the median, whereas the distribution of detector response values for BUFFER1 are heavily skewed, and thus certainly not normally distributed.

We study the detector response (intensity output) for SELDI under varying input conditions, creating a detector response curve as follows. For each fixed time (mass) point across spectra from BUFFER1 in the mass focused region $[3kDa, 30kDa]$, we estimate the mean intensity observed and the corresponding variance, with the same repeated for BUFFER2. These are displayed as a scatter plot in Figure 9 along with the best fit quadratic curve.

Observing Figure 9 we see

1. Intensity fluctuation/variance increases monotonically with the mean.

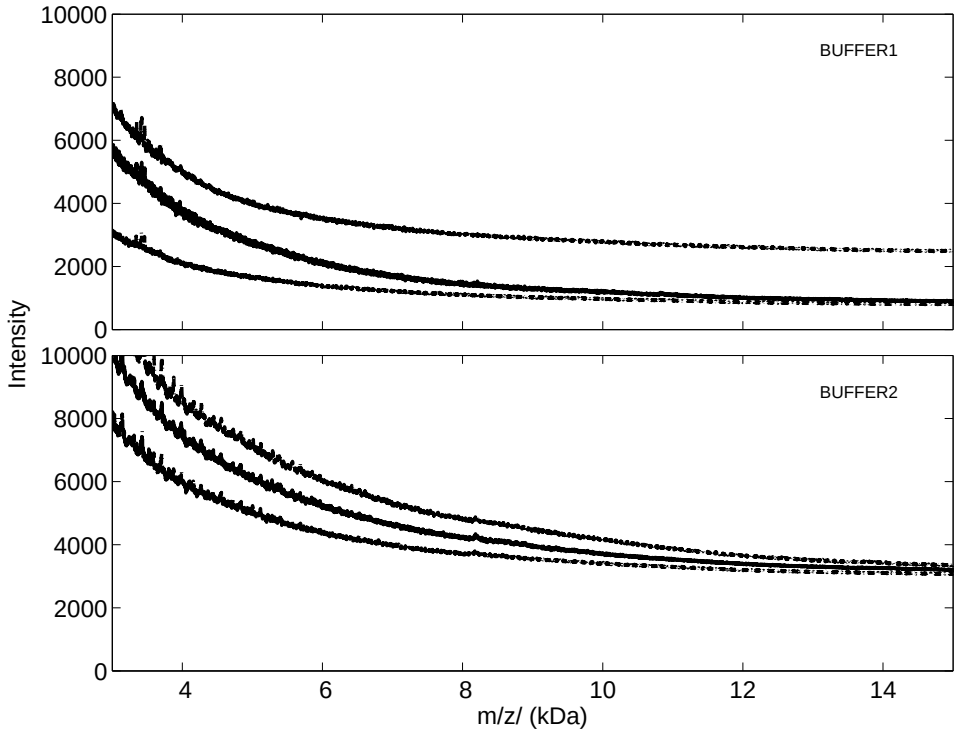

Figure 8: Quantile spectrum visualizations for all 183/114 spectra from BUFFER1/BUFFER2 datasets respectively. The middle, upper, and lower spectra are the 50% (median), 75%, and 25% quantile spectra respectively, calculated pointwise for each mass point. The results show that different machine settings give rise to different statistical behavior of the intensity values registered at the detector. Preprocessing techniques should be able to adapt to this varying behavior.

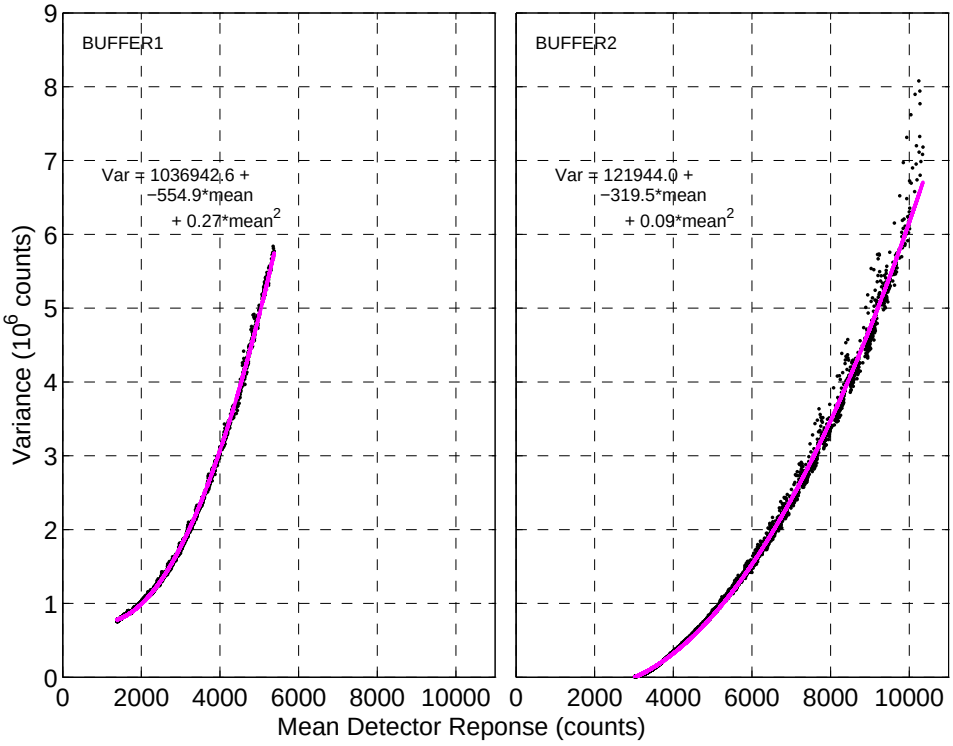

Figure 9: SELDI detector response curves. For repeated experiments under homogeneous machine settings, the variance in intensities observed is shown to be quadratic in the mean intensity observed. Thus, peaks occurring in areas of the spectrum affected near the baseline will be more noisy and more difficult to detect. Most algorithms for preprocessing SELDI data assume constant variance, independent of signal intensity. The detector response curve is shown to be dependent on machine settings, as it is different for BUFFER1 and BUFFER2.

2. The variance of the detector response is a quadratic function of the mean, to a very good approximation

3. The detector response curves for BUFFER1 and BUFFER2 are quite different, and thus are dependent on the machine settings.

The detector response statistics thus exhibit a quadratic variance function. Briefly, a random variable X is said to have a quadratic variance function (QVF) if

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2, \tag{11}$$

with μ being the mean of X, $V(\mu)$ the variance, and v_0, v_1, v_2 constants, some of which may be zero.

From these observations, summarized in Figures 8 and 9, it seems unlikely that an algorithm optimized for BUFFER1 would work well on BUFFER2 and vice versa. Further, neither a homoscedastic approach (e.g. - standard wavelet shrinkage [33]) or a simple heteroscedastic approach (e.g. - Poisson regression formulation [34]) to preprocessing the data is likely to be sufficient.

3.3.2 Data for evaluating preprocessing algorithms

We have generated two new datasets for evaluating preprocessing algorithms in order to improve upon purely simulation-based datasets used in previous comparison studies [24,39]. A good comparison dataset should have the following properties (discussed previously in [39]):

1. Exact protein content is known (and thus expectation of where "true" peaks will appear)

2. Analyzed sample is complex containing many proteins/peaks

3. Noise and baseline characteristics should be as close to those of real SELDI data as possible.

If one uses simulated data [24,39,82], complete control can be attained over requirements 1) and 2) at the expense of having noise/baseline characteristics that are overly ideal. If one

uses purely real data, the noise, baseline, and artifacts that arise in actual experiments are present. However, this usually accompanies the trade-off of either not knowing the exact protein content (e.g.- complex serum data) or an overly simplified scenario (e.g. - spike-in data).

We combine the advantages of purely simulated and real data by introducing the notion of a hybrid spectrum. To generate a hybrid spectrum, we use an implementation of the SimSpec 2.1 SELDI simulator [18, 82][1] to generate a "clean" SELDI spectrum, shown at the top of Figure 10. This gives an accurate peak shape characteristic as would be seen in low resolution SELDI/MALDI for given mass and ion abundance values, without any electronic noise or baseline present. We then select one of our buffer + matrix spectra (from either BUFFER1 or BUFFER2) and add the two together to produce the hybrid spectrum shown at the bottom of Figure 10. Thus, in a hybrid spectrum we know the *exact* virtual protein content specified to the simulator *a priori* while maintaining *exactly* the same noise, baseline, and other artifacts one encounters with real SELDI data.

Further details on the hybrid spectra can be found in the Methods section and in additional file 1: supplement.pdf. The collection of hybrid spectra under different operating conditions results in test sets, denoted HYBRID1 and HYBRID2, with each test set containing thirty datasets of fifty hybrid spectra each. The mean performance of a preprocessing algorithm on HYBRID1 and HYBRID2 can be interpreted as the expected performance of the preprocessing approach in each separate operating condition in a repeated experiment or sampling from a homogeneous population (e.g. - cancer group or control group).

3.3.3 New preprocessing algorithms for SELDI

We have developed a set of MATLAB® scripts for preprocessing SELDI spectra named LibSELDI. For information on how to obtain LibSELDI and the associated scripts used to produce the figures in this paper, see additional file 1: supplement.pdf. We compare our preprocessing package to the MassSpecWavelet package from the Bioconductor project [35]. MassSpecWavelet has been established as one of the best approaches in terms of peak finding

[1] http://bioinformatics.mdanderson.org/Software/Cromwell/simspec.zip

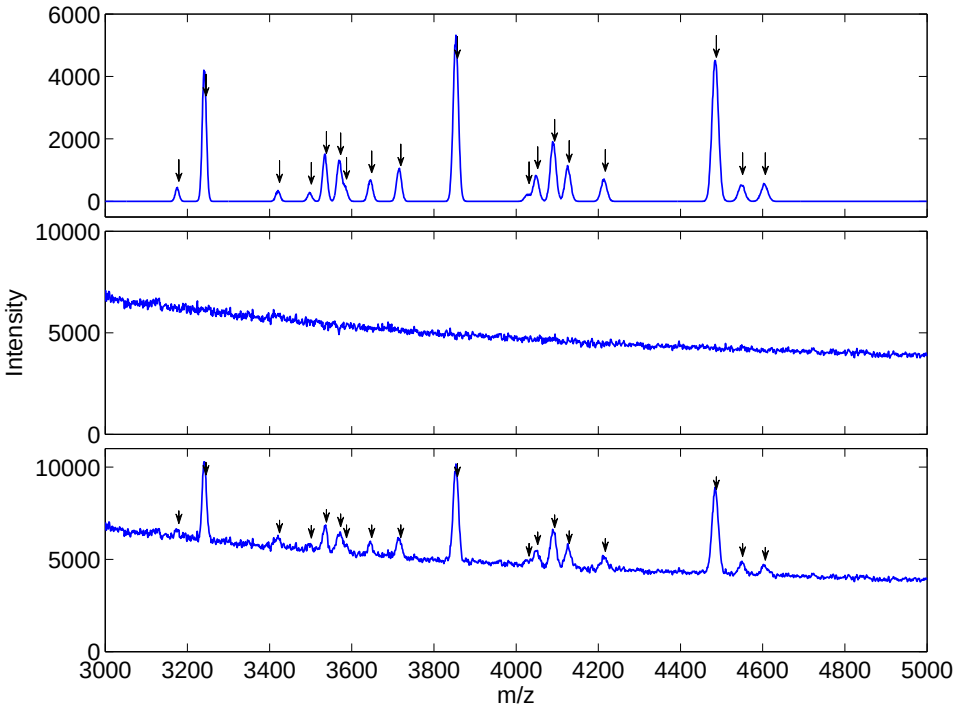

Figure 10: Construction of hybrid spectrum for testing preprocessing algorithms. (top) Clean, pure protein component spectrum with no noise and no baseline simulated using SimSpec 2.1 MALDI/SELDI simulation engine. Arrows over peaks show the m/z values of the virtual proteins. (middle) Buffer+matrix spectrum generated in a SELDI PBS IIc, representing noise, baseline, and artifacts that are typically seen. (bottom) Final hybrid spectrum, consisting of the sum of simulated and real components. Hybrid spectra have the advantage of having diverse signal components (150 virtual proteins) with *exact* knowledge of the virtual proteins while retaining the true noise and baseline characteristics from real SELDI data.

in recent comparison studies [24, 39], and has been downloaded > 6000 times in the past two years as of March 2010[2]. Both packages have the advantage of having only one main user-adjusted parameter.

In order to compare the performance of each preprocessing program, we generate operating characteristic curves (OC curves) [39,79], one for each of the 30 datasets of HYBRID1 and HYBRID2, by varying the Peak Area threshold (LibSELDI) and signal-to-noise ratio threshold (Snr.Th in MassSpecWavelet) parameters in the programs. Code snippets showing how MassSpecWavelet was tested can be found in additional file 1: supplement.pdf. This allows us to understand the trade-offs between false discovery rate (FDR) and sensitivity (TPR) achieved by each algorithm. The results for both the HYBRID1 and HYBRID2 collections are shown in Figure 11, where we have plotted the fdr-axis in log scale to emphasize the low FDR region which is usually of most interest in biomarker discovery applications. Note that, since both HYBRID1 and HYBRID2 are collections of datasets representing repeated trials (or equivalently a homogeneous population), the OC curves we show in Figure 11 are the mean OC curves across the 30 datasets for each.

The results show that LibSELDI tends to have a considerable advantage in the low FDR region, while MassSpecWavelet tends to have higher sensitivity for $FDR > 25\%$. One way to summarize the performance of the algorithms is using the area under the OC curve for the FDR region of interest. We compute two area under the curve values, PAUC [39] (calculated for $FDR \in [0, 50\%]$), and PAUC25 (calculated for $FDR \in [0, 25\%]$). The results are shown in Table 7, where we have normalized each score separately so that a perfect PAUC25 (likewise, PAUC50) score is 100.

In Figure 12, we show the specific operating characteristics for LibSELDI and MassSpecWavelet for Dataset 2 of HYBRID1. While both algorithms perform well, LibSELDI resolves more than 90 proteins correctly before making a mistake. Since operating characteristics show false discovery rate along the x-axis rather than false positive rate (as in the traditional ROC curves), they tend to penalize more when false predictions are made with very few true proteins found. Indeed, in this case MassSpecWavelet got its first protein prediction

[2] `http://bioconductor.org/packages/stats/bioc/MassSpecWavelet.html`

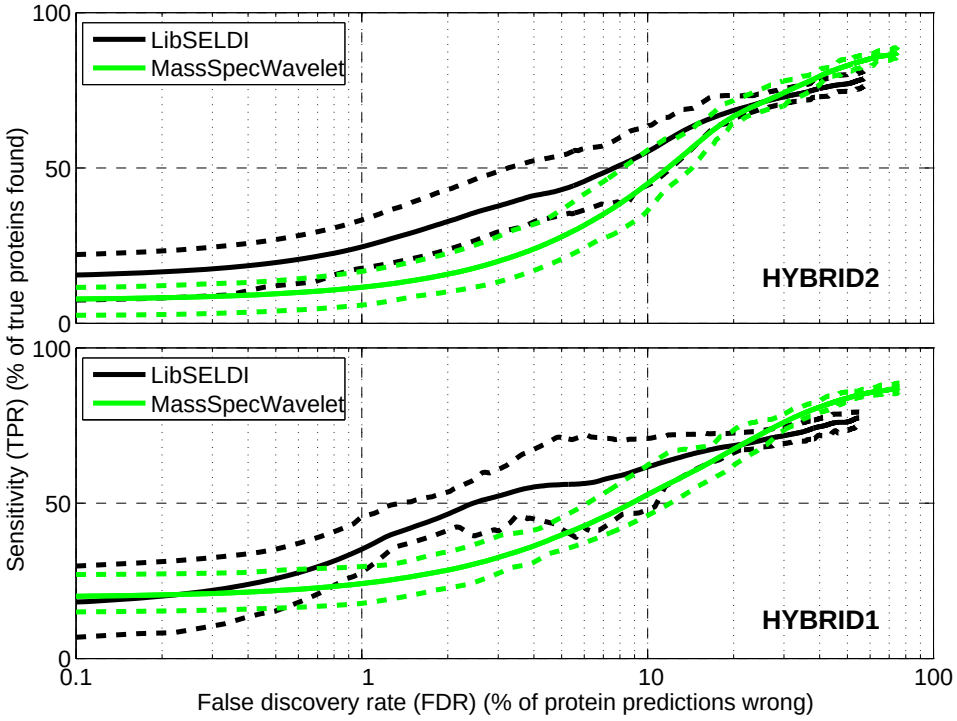

Figure 11: Trade off between sensitivity and false discovery rate for LibSELDI and MassSpecWavelet. Average loess-smoothed operating characteristics show the trade-offs between sensitivity (TPR) and false discovery rate (FDR) for HYBRID1 and HYBRID2. The mean loess-smoothed curve is indicated by the solid line, while the upper and lower dashed lines indicate the 75% and 25% quartile curves. The FDR axis is shown in log-scale to emphasize lower FDR values. LibSELDI demonstrates superior sensitivity compared to MassSpecWavelet on both datasets for FDR values less than about 25%. MassSpecWavelet has the advantage for FDR values greater than 25%.

Table 7: Area under the operating characteristic comparison. Area under the operating characteristic curve in a range of false discovery rate values of interest is a useful way to compare peak prediction performance. We show two partial area under the curve metrics, calculated in the range $FDR \in [0, 50\%]$ (PAUC) and $FDR \in [0, 25\%]$ (PAUC25). PAUC is more of overall measure of peak prediction potential, while PAUC25 focuses on measuring performance at low FDR. The number shown is the average (standard error) calculated from the 50 operating curves from HYBRID1 and HYBRID2. LibSELDI shows particularly appealing PAUC25 performance

Algorithm/Dataset	PAUC25	PAUC
LibSELDI/ HYBRID1	58.8% (9.9)	66.1% (7.1)
MassSpecWavelet/ HYBRID1	50.8% (8.5)	64.9% (5.9)
LibSELDI/ HYBRID2	53.2% (8.7)	64.1% (6.1)
MassSpecWavelet/ HYBRID2	45.4% (9.5)	61.3% (6.9)

correct but its second prediction wrong, leading to the point at FDR=50%, TPR=7%. Thus, operating characteristics with false discovery rate along the x-axis enforce the principle of conservative decision making, rewarding approaches that are successful with their initial large threshold (conservative) predictions and penalizing those that make mistakes early.

At FDR values greater than 30%, MassSpecWavelet outperforms LibSELDI. However, this is at the expense of generally more promiscuous predictions, since MassSpecWavelet generates 586 potential protein predictions compared to 250 for LibSELDI.

3.4 Discussion

We posit that the detector response is a member of the Natural Exponential Family with Quadratic Variance Function (NEF-QVF), which is a proper subset of the exponential family of distributions [81]. Figures 8 and 9 show that assuming the detector response takes the form of a specific distribution is impractical, but that the detector response $V(\mu)$ has a QVF. The NEF-QVF family of distributions occur often in practice and have the following useful properties, characterized by Morris [81]:

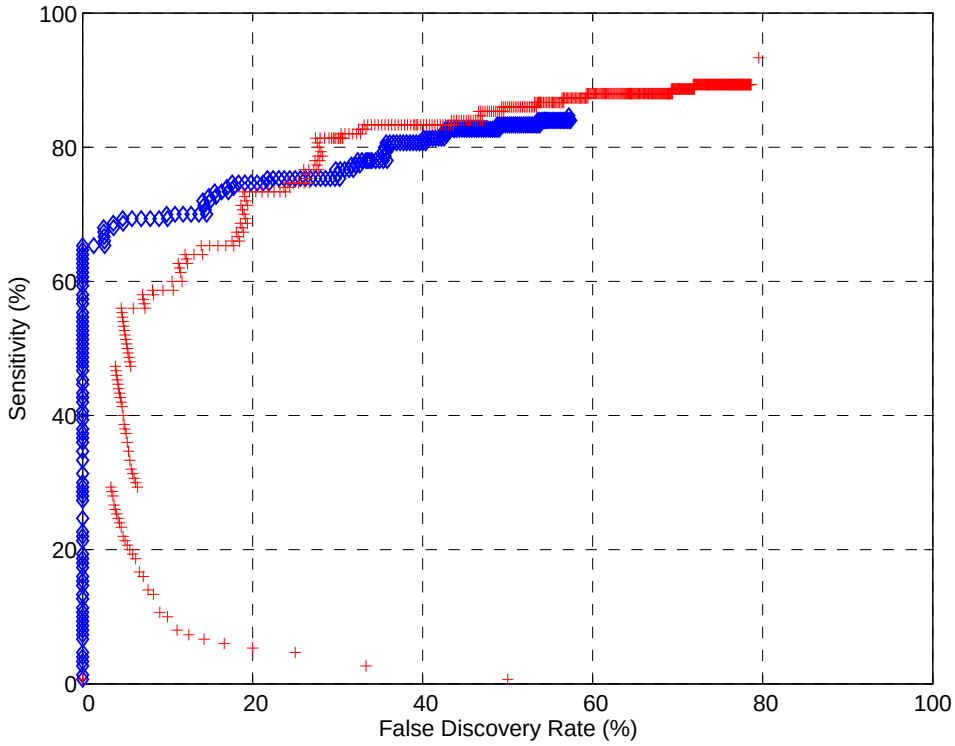

Figure 12: Example operating characteristic. Operating points shown summarize the performance of LibSELDI and MassSpecWavelet on Dataset 2 of HYBRID1 for many different parameter choices. Each blue diamond is the (FDR, TPR) observed for a single choice of Peak Area threshold for LibSELDI, while each red plus symbol shows the result of a single Snr.Th parameter choice for MassSpecWavelet. For this particular example, LibSELDI finds more than 90 true proteins before making a mistake. At high FDR conditions, MassSpecWavelet resolves close to 90% of proteins compared to about 85% for LibSELDI.

1. If a random variable $X \in$ NEF-QVF, it is completely specified by its variance function $V(\mu)$

2. If $X \in$ NEF-QVF, a, b constants then $aX + b$ is also NEF-QVF

3. **Additivity:** If $X_1, X_2 \in$ NEF-QVF, then $X_1 + X_2$ is NEF-QVF

4. Linear combinations of normal, Poisson, gamma, binomial, negative binomial, and generalized hyperbolic secant distributed random variables generate all possible distributions in the NEF-QVF family.

There are some physical reasons as to why the NEF-QVF assumption could be reasonable as well. Some plausible justifications for the first two terms in Eq. (11) are:

1. **Constant Term:** This is possibly due to thermal noise (additive Gaussian noise) which is common to all electronic measurement devices [111]

2. **Linear Term:** The ability to detect an ion in a multiple stage electron multiplier, a common type of detector in MALDI-like instruments, is described by compound Poisson statistics [31].

The existence of a plausible physical explanation for the quadratic variance term remains an open question. However its effect is measured in both BUFFER1 and BUFFER2 and cannot be neglected. While the QVF model explains the data well in the mass focused region between 3 and 30 kDa, it is likely to break down at lower masses around 2-2.5 kDA where the baseline reaches a maximum. In this region the detector often saturates, introducing a non-linearity into the data that we have not accounted for.

The success of our univariate model for SELDI may indicate that we have selected the most important feature to consider in the preprocessing of the data: namely, the fluctuations in the response of the ion detector subject to different inputs. The analysis of expression values of preprocessed data, on the other hand, requires multivariate methods as there are significant statistical dependencies between the peak heights corresponding to proteins that may be interacting. While these correlations are important in the analysis performed after

the data is preprocessed, our results indicated it may be safe to ignore them during the preprocessing.

It is entirely possible that the quadratic variance model could be applicable to other similar technologies such as MALDI and newer SELDI mass spectrometers. This, however, has not been confirmed.

Having buffer only spectra allows one to estimate the parameters of the detector response curve. Knowledge of the detector response curve enables us to apply the modified Antoniadis-Sapatinas denoising scheme described in the methods. Using this approach in our LibSELDI package yields excellent peak detection performance. We have proved this concept on HYBRID1 and HYBRID2 by estimating the QVF parameters of (11) using the buffer-only spectra that were randomly selected from BUFFER1 and BUFFER2 respectively. This implies that spots on SELDI chips should be reserved for buffer-only spectra. Thus, the trade-off for using our approach is increased cost in terms of the number of chips one must use. The modified Antoniadis-Sapatinas denoising is computationally intensive as well, taking approximately seven minutes per spectrum on a high-end workstation.

We argue that some of the cost is recovered by the potential for adaptive and accurate preprocessing, but not all. It may be possible to use QC and/or calibration samples to estimate the QVF as well rather than buffer-only spots. However, this would add in some additional variation due to the nature of the medium (serum, plasma, etc).

While LibSELDI outperforms MassSpecWavelet on the HYBRID1 and HYBRID2 test sets, the applicability of this comparison and of these results to purely real data remains an open question. There is some basic biological variability modeled in our test sets (see description in Supplement of [39]). However, data from complex biological samples such as serum or plasma likely contains more biological variation and artifacts than we have modeled in HYBRID1 and HYBRID2. The investigation of how biological variation affects the model in QC samples is a work in progress.

In addition to achieving a better mean OC curve at lower FDR values, LibSELDI consistently predicts fewer peaks than MassSpecWavelet, leading to protein predictions closer to the true number of proteins in the data, as shown in Figure 13. This is further evidence

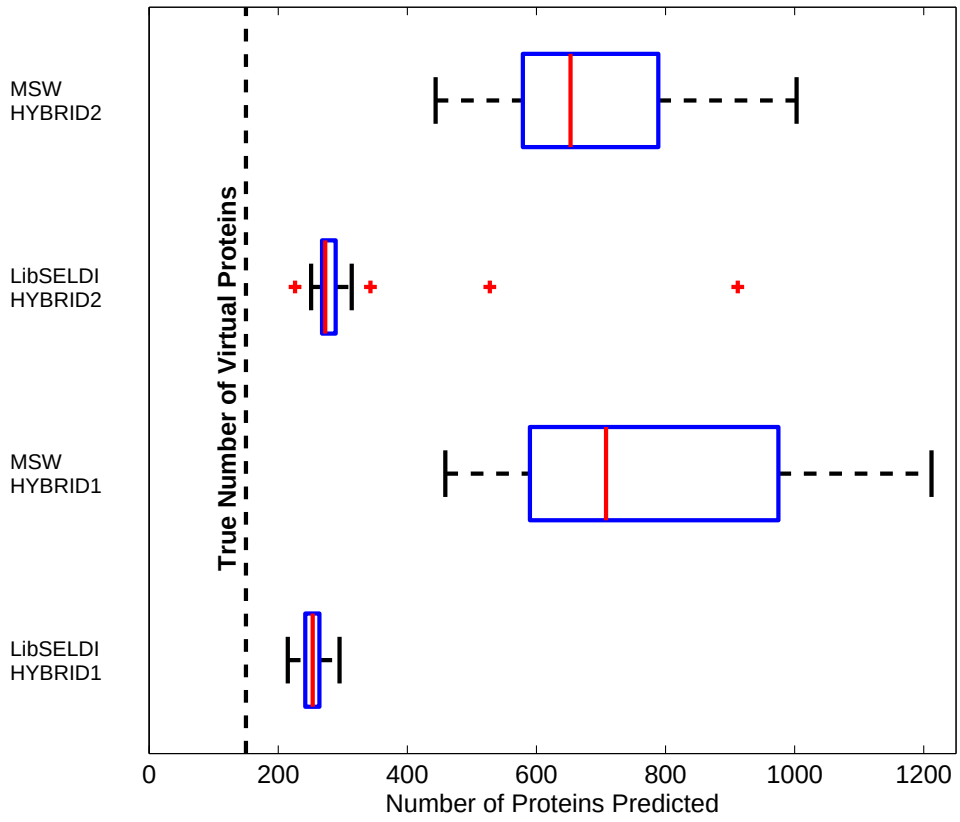

Figure 13: Efficiency of peak/protein predictions. We show boxplots summarize the number of peaks predicted for each program in the mean spectrum of each dataset from HYBRID1 and HYBRID2 before thresholding. LibSELDI consistently predicts around 250 peaks, while MassSpecWavelet predicts more than 600 peaks consistently. MassSpecWavelet's more promiscuous predictions lead to high sensitivity at the expensive of higher false discovery rate performance. LibSELDI's peak predictions are reproducibly closer to the true number of virtual proteins, 150 of them, present in each dataset.

that the adaptive modified Antoniadis-Sapatinas denoising approach using the NEF-QVF model for the detector response is smoothing the spectra by close to the right amount.

3.5 Conclusions

We have shown that the variance of the intensity of a SELDI spectrum is quadratic in the mean signal strength. We further make the flexible assumption that the underlying distribution of the intensities is from a natural exponential family. From this point of view, we use a modified Antoniadis-Sapatinas wavelet shrinkage approach for denoising SELDI spectra. With this method at the core of our LibSELDI program for preprocessing SELDI

data, we demonstrate excellent sensitivity at low false discovery rates. For applications that can tolerate higher false discovery rates, the MassSpecWavelet algorithm performs better in this region.

Our work has implications in the design of SELDI experiments. Namely, the modified Antoniadis-Sapatinas denoising technique performs well but requires an estimate of the quadratic variance function (QVF) describing the SELDI detector. This, in turn, is affected by machine settings. We have used buffer-only spectra to estimate the QVF. Thus, buffer-only spots could be interlaced on chips. We are investigating less expensive ways to estimate the QVF in future work.

3.6 Methods

3.6.1 Protocol for generating buffer-only spectra

Buffer-only spectra were generated by interspersing buffer only samples with protein samples from subjects (e.g. serum samples) and with pooled subject samples (for quality control) on the same chip. The buffer-only samples were spotted with wash buffer that was either PBS (phosphate buffered saline with various concentrations of phosphate and NaCl) based or acetonitrile + TFA (triflouroacetic acid) based, as manufacturer recommended per chip type. These buffer only samples were processed with the same washing steps as the subject samples, as described in [96], and then SPA matrix was applied to all spots.

The samples were analyzed with the Protein Biological System II-c^{TM} SELDI mass spectrometer (Bio-Rad Laboratories, Inc., Hercules, CA). The machine settings (e.g. laser intensity, detector sensitivity) and precise washing steps varied from buffer only spot to buffer only spot, and were generally different between BUFFER1 and BUFFER2. Note especially that laser intensities were generally higher for BUFFER2 than for BUFFER1. A detailed list of machine settings is given in the supplement.

3.6.2 Hybrid data

Calculating performance statistics for comparison of MassSpecWavelet and LibSELDI requires a large number of spectra emulating an experiment that was repeated many times. To generate the HYBRID1 dataset, we combine each clean spectrum with one buffer+matrix

spectrum from BUFFER1, and similarly we form HYBRID2 from BUFFER2 by combining those spectra with the same clean spectra.

A basic model of repetitive experiments for SELDI is available with SimSpec 2.1 that takes into account fluctuations in protein concentrations, m/z values, and prevalence in the data. Using the SimSpec 2.1 model developed at the MD Anderson Cancer Center [18, 82], we generate 30 datasets containing 50 clean (noise and matrix-free) spectra each. Each dataset consists of 150 virtual proteins and each spectrum within the given dataset contains a proper subset of these proteins with fluctuating parameters according to the model described in [82] and its supplement. The goal for the preprocessing programs in our performance evaluation is to reconstruct the master list of 150 virtual proteins characterizing the dataset. Repeated across all 30 datasets, we can calculate useful performance statistics. The properties of the 150 virtual proteins themselves are drawn from a prior distribution that was estimated from real data. See [82], or alternatively, the description in the supplement of [39].

We use sampling to overcome the limitation of having much fewer spectra in BUFFER1 and BUFFER2 than we have clean spectra in preparation for testing the algorithms. In principle the best way to construct the hybrid test sets would be to have one unique spectrum in BUFFER1 (likewise BUFFER2) for each spectrum in our clean protein-only set. However, this would require 1500 buffer+matrix runs to be performed for both BUFFER1 and BUFFER2, an impractical amount of blank chips to run. Sampling from BUFFER1 (BUFFER2) provides a cost effective way to introduce variation in the noise/matrix characteristics between the datasets in HYBRID1 (HYBRID2).

3.6.3 Preprocessing the spectra

First we consider a model for a single SELDI spectrum, $X(t)$. We observe $X(t)$, a random process, on a discrete time grid $t_1, \ldots, t_m$, where $X(t)$ represents the intensity of the raw SELDI spectrum observed at time (equivalently mass) point t. For all t, we assume that $X(t)$ is distributed according to a natural exponential family (NEF) with quadratic variance function (QVF) equal to $V(\mu(t))$ as in Eq. (11). The variance function $V(\mu)$ completely

characterizes the NEF-QVF family. The goal of preprocessing in SELDI is to estimate $\mu(t)$, the expectation of $X(t)$, which is the signal corresponding to ions that hit the detector. With a good estimate of $\mu(t)$, extracting peaks and estimating protein m/z values in a dataset is relatively straightforward.

As a side note we point out that a SELDI spectrum is actually a sum of single shot spectra. However, the additivity property of the NEF-QVF family guarantees the sum is NEF-QVF provided that the single-shot spectra are NEF-QVF, agreeing with our detector response model and experimental observations.

3.6.3.1 Multiple spectra considerations

Rather than observe a single spectrum, the typical biomarker discovery approach is to generate at least one spectrum for each of n samples from an approximately homogeneous population. For example, one homogeneous population may be a group of early stage prostate cancer patients matched for age, race, etc. Assuming the samples are run on the same SELDI machine with the same operating conditions, we have

$$X_1(t), \ldots, X_n(t) \propto \text{NEF-QVF}\left(V(\mu(t))\right). \tag{12}$$

Our assumption that all n patients have the same underlying $\mu(t)$ is equivalent to assuming that the underlying biological condition being observed in each patient is approximately the same. Thus, we wish to estimate the underlying commonality $\mu(t)$ related to the biology of their condition expressed through the SELDI signal. We can mitigate some of the effects of the QVF by forming the mean spectrum (first introduced by [82]).

$$X_\bullet(t) = \frac{1}{n}\sum_{k=1}^{n} X_k(t). \tag{13}$$

It is straightforward to show that

$$E\left\{X_\bullet(t)\right\} = \mu(t) \tag{14}$$

$$\text{Var}(X_\bullet(t)) = \frac{1}{n}V(\mu(t)). \tag{15}$$

Thus, the mean spectrum concept is valuable under the assumptions of the NEF-QVF model as well.

3.6.3.2 Modified Antoniadis-Sapatinas denoising

We now discuss estimation of $\mu(t)$ from the mean spectrum (13). Since the $X_k(t)$ are sampled on a discrete time grid (and thus $X_\bullet$), we introduce vector notation

$$\mathbf{x}_\bullet = [X_\bullet(t_1), \ldots, X_\bullet(t_m)]'$$
$$\boldsymbol{\mu} = [\mu(t_1), \ldots, \mu(t_m)]'.$$

For any estimate $\hat{\boldsymbol{\mu}}(\mathbf{x}_\bullet)$ of $\boldsymbol{\mu}$, we measure its fitness using the mean-squared-error (MSE)

$$MSE(\hat{\boldsymbol{\mu}}(\mathbf{x}_\bullet), \boldsymbol{\mu}) = E\left\{\|\hat{\boldsymbol{\mu}}(\mathbf{x}_\bullet) - \boldsymbol{\mu}\|^2\right\}. \tag{16}$$

Antoniadis and Sapatinas proposed a wavelet shrinkage scheme to solve for $\hat{\mu}$ in (16) in the context of NEF-QVF regression [3]. We summarize their main results. For our denoising, we use the orthogonal discrete wavelet transform with respect to the Symmlet 8 basis [25]. The transform can be represented by an $m \times m$ orthogonal matrix W,

$$\mathbf{w} = W\mathbf{x}_\bullet. \tag{17}$$

Let $\mathbf{h}$ be a length m vector with entries taking values between 0 and 1. Let $H = \text{diag}(\mathbf{h})$ be the $m \times m$ matrix defined by placing the entries of $\mathbf{h}$ along the main diagonal, all other entries 0. The class of estimators for $\hat{\boldsymbol{\mu}}(\mathbf{x}_\bullet)$ considered by [3] take the form

$$\begin{aligned}\hat{\boldsymbol{\mu}}(\mathbf{x}_\bullet) &= W'H\mathbf{w} \\ &= W'HW\mathbf{x}_\bullet.\end{aligned} \tag{18}$$

This is the typical wavelet denoising scenario where each wavelet coefficient is left alone or shrunk towards zero according to some criterion, and is completely defined by the vector $\mathbf{h}$. Antoniadis and Sapatinas showed that a good estimator for data from the NEF-QVF family is given by choosing

$$\hat{\mathbf{h}}(i) = \frac{\left[\mathbf{w}(i)^2 - \hat{\boldsymbol{\sigma}}^2(i)\right]_+}{\mathbf{w}(i)^2}, \quad i = 1, \ldots, m$$

$$[z]_+ = \begin{cases} z, & z \geq 0 \\ 0, & z < 0. \end{cases}$$

The term $\hat{\boldsymbol{\sigma}}^2$ is estimated as

$$\hat{\boldsymbol{\sigma}}^2 = \frac{1}{1+v_2}(W \cdot W)V(\mathbf{x}_\bullet). \tag{19}$$

Where $V(\mathbf{x}_\bullet)$ is the vector constructed by applying the QVF from (11) to each term of $\mathbf{x}_\bullet$. $(W \cdot W)$ is the matrix whose i,j element is the square of the i,j element of W. The parameters v_0, v_1, v_2 in (11) are measured from the buffer-only spectra, as described in the Results and Discussion section.

We make an intuitive modification to (19)

$$\widetilde{\boldsymbol{\sigma}}^2 = \frac{1}{1+v_2}(W \cdot W)V^\dagger(\mathbf{x}_\bullet)$$
$$V^\dagger(\mathbf{x}_\bullet(i)) = \max\left\{V(\mathbf{x}_\bullet(i)), v_0\right\}.$$

Thus our modified Antoniadis and Sapatinas estimator $\widetilde{\mathbf{h}}$ uses $\widetilde{\boldsymbol{\sigma}}^2$ in (18) rather than $\hat{\boldsymbol{\sigma}}^2$. The modification was introduced to account for cases when (19) may underestimate the noise when low amounts of observed signal are detected. Define

$$\widetilde{\mathbf{h}} = \frac{\left[\mathbf{w}(i)^2 - \widetilde{\boldsymbol{\sigma}}^2(i)\right]_+}{\mathbf{w}(i)^2}$$
$$\widetilde{H} = \mathrm{diag}(\widetilde{\mathbf{h}}).$$

Then, our modified Antoniadis-Sapatinas estimate of $\boldsymbol{\mu}$ is defined as

$$\widetilde{\boldsymbol{\mu}} = W'\widetilde{H}W\mathbf{x}_\bullet. \tag{20}$$

3.6.3.3 Peak detection/baseline removal

We consolidate the two preprocessing steps of baseline removal and peak detection typically performed separately into a single step as follows. We assume that the underlying $\mu(t)$ shown in (14) is the superposition of protein ions, $s(t)$, and energy-absorbing matrix ions, $b(t)$ striking the detector. It is well known that the distribution of the isotopes in our analyte of interest gives rise to a roughly Gaussian peak shape. Thus, we propose

$$\mu(t) = s(t) + b(t) \tag{21}$$
$$s(t) = \sum_j a_j \mathcal{G}_{3\sigma_j}(t_j, \sigma_j) \tag{22}$$

where, $\mathcal{G}_\alpha(t_j, \sigma_j)$ denotes a Gaussian kernel function centered at t_j with standard deviation σ_j and zero outside the interval $[t_j - \alpha, t_j + \alpha]$.

Typically, $s(t)$ is very sparse in the sense that it is mostly zero over the domain of the observed signal. Therefore, the local minima of our estimated baseline + noise signal $\widetilde{\boldsymbol{\mu}}$ are points we may assume touch the baseline. From this point of view, once we have detected all the local minima in $\widetilde{\boldsymbol{\mu}}$, the baseline curve estimation problem reduces to an interpolation problem amongst these points. We have found through experimentation that piecewise cubic Hermite interpolating polynomials [45] are excellent interpolation functions.

The minima and maxima in $\widetilde{\boldsymbol{\mu}}$ are found in one pass using the extrema function downloadable from MATLAB® central file exchange. The maxima are the peaks in the mean spectrum potentially indicating proteins represented in our sample population while the minima correspond to samples from the baseline signal.

Each detected peak is quantified using peak area and a threshold is chosen based on the peak area measurement to generate the final prediction set.

3.6.4 Operating characteristics

The peaks we detect in $\widetilde{\boldsymbol{\mu}}$ represent the initial set from which we choose our final estimates of proteins that are active in the population of interest. The choice of final estimate is accomplished using a peak area threshold (LibSELDI) or signal-to-noise ratio measurement (Snr.Th in MassSpecWavelet). From each prediction, we calculate the observed false discovery rate (FDR) and true positive rate (TPR, also called sensitivity)

$$FDR = \frac{FP}{FP + TP} \tag{23}$$

$$TPR = \frac{TP}{TP + FN}. \tag{24}$$

Where TP (the number of true positives) is the number of the 150 virtual protein m/z values having at least one predicted m/z value within 0.3% relative error. The FP is defined as the number of predicted m/z values not within 0.3% of any of the 150 virtual protein m/z values for this dataset. Similarly, FN is the number of the 150 virtual protein values without any predicted m/z value within 0.3% relative error.

For each dataset, a curve is fit to the operating points. Each operating curve is averaged to produce a mean operating characteristic, as shown in Figure 11. From this curve, the calculation of the area-under-the curve is straightforward. For more details, see sections 2.2 and 2.2.1 of [39].

CHAPTER IV

EXPLAINING REPRODUCIBILITY OF PEAKS IN SELDI MASS SPECTROMETRY: THE QUADRATIC VARIANCE MODEL

4.1 Introduction

The reproducibility of peaks in surface-enhanced laser desorption/ionization time of flight mass spectrometry (SELDI) has been problematic. In 2002, Petricoin developed an approach for the early detection of ovarian cancer based on SELDI [90]. Further scrutiny of this study revealed artifacts in the data that biased the results given by Petricoin [5, 103]. This led to several important papers studying experimental pre-analytic and analytic factors affecting reproducibility [77, 78, 100, 109]. Recently, several studies have been performed studying post-analytic factors of reproducibility, namely, the preprocessing of the data [24, 39, 80, 117]. These studies indicate clearly that the choice of preprocessing algorithms leads to significantly different results with respect to the quality of the peaks found in the data. Continuing concerns about reproducibility are highlighted in the recent article by Wei *et al* [118].

One factor contributing to our poor understanding of reproducibility for SELDI is the lack of bottom-up models characterizing the measurement process. By default machine settings, a SELDI spectrum is the result of pooling/summing numerous single-shot spectra. Sköld *et al* studied the acquisition of single shot spectra and proposed a statistical framework for pooling the single shot spectra [102]. They introduced an expectation-maximization algorithm for combining the spectra that results in improved peak heights in the pooled spectrum. Malyarenko *et al* introduced a charge-decay model for the baseline in a SELDI spectrum and used time-series methods for the common preprocessing tasks. Recently, Emanuele and Gurbaxani have proposed a quadratic variance model for the response of the detector to buffer-only plus matrix sample runs, which leads to preprocessing methods showing improved performance [40].

Cervical mucous is an important biological medium to understand in order to increase the chances of finding biomarkers for the early detection of cervical cancer. Recently there have been efforts to characterize the cervical mucous proteome (e.g. [87]). In this paper, we test the sufficiency of the quadratic variance model for explaining SELDI spectra generated from pooled cervical mucous QC data and finding reproducible peaks.

4.2 Materials and Methods

4.2.1 Cervical Mucous and Patients

Cervical mucous samples were collected from women attending coloposcopy clinics as described in [87].

4.2.2 SELDI mass spectrometry

Pooled cervical mucous was spotted on chips intermittently as part of a QC step in the experiment design. Patient samples were also spotted on the same chips although we do not use the patient cervical mucous samples in this study.

Mass spectrometry analysis was performed using the Protein Biology System IIc SELDI mass spectrometer (Bio-Rad Laboratories, Hercules, CA). Mass focusing was optimized for the 3 kDa to 30 kDa region. Data collection from start to finish took 2 weeks. We collected the spectrum from pooled mucous QC samples for the low laser setting, CM10 array, for analysis in this study. In total, there were 36 spectra. Five of the spectra had already been through one freeze/thaw cycle before analysis. These spectra were removed leaving us with a final dataset of 31 pooled-mucous QC spectra.

4.2.3 Quadratic variance model

We fit a quadratic model to the variance of the data by measuring the mean and variance from hand selected regions of the QC spectra where peaks are visibly absent in all of the spectra. The quadratic variance model implies that the mean μ and variance $V(\mu)$ have the relationship

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2. \tag{25}$$

Emanuele and Gurbaxani have previously shown this is a good model for measurements of the resulting spectra from buffer washed, matrix-only spectra containing no biological signal or protein content [40]. In principle, the measurement of the data in between peaks should be a sufficient approximation to this as long as we take it over a large enough mean intensity range to enable good curve fitting.

The quadratic variance model can be used to make testable predictions about how peaks should behave in the spectra of a SELDI experiment. One subtle aspect of Eq. (25) is that it predicts what the CV of our measurements should be,

$$\begin{aligned} CV\% &= 100 \cdot \frac{\sigma}{|\mu|} = 100 \cdot \sqrt{\frac{V(\mu)}{\mu^2}} \\ &= 100 \cdot \sqrt{v_0\mu^{-2} + v_1\mu^{-1} + v_2} \qquad (26) \\ &\approx 100 \cdot \sqrt{v_2} \quad (\mu \text{ large}). \qquad (27) \end{aligned}$$

We summarize the deductions/predictions made by the quadratic variance model that will be tested in this paper:

1. Peak heights across spectra corresponding to the same underlying biology should have mean heights and variances consistent with the estimated quadratic variance function of (25)

2. The measured CV values of peak heights should be consistent with (26)

3. For very large peaks, the CV of peak heights should be approximated by (27)

4. An algorithm that uses the quadratic variance nature of the measurements (LibSELDI [40]) should find more reproducible peaks than Ciphergen Express [46].

4.2.4 Preprocessing with LibSELDI

The LibSELDI preprocessing package being developed in MATLAB®(The Mathworks, Natick, MA) is the first to take into account a quadratic variance form of the measurement error. The details of the algorithms used with LibSELDI have been described previously [40]. We use this technique to process the data adhering to the following protocols:

- A single quadratic variance function (QVF) is estimated representing all 31 QC spectra
- The QVF is estimated according to the procedure described in Sec. 4.2.3
- Preprocessing is performed on each spectrum individually rather than the mean spectrum as was done in [40].

4.2.5 Preprocessing with Ciphergen

Ciphergen Express is the most popular preprocessing suite for working with SELDI data. The precise procedure used to preprocess the data and produce peak predictions with Ciphergen is described in [86].

4.2.6 Peak matching algorithm

We propose an algorithm for matching peaks with similar m/z values in different spectra that are assumed to be from the same underlying protein. A useful formulation of this problem is using graph theory, where peaks are represented by vertices and two vertices have an edge between them if their m/z values are within a specified tolerance. A clique is defined as a set of vertices for which all pairs of vertices contain an edge. In other words, a set of peaks that are all sufficiently close to each other in m/z value to be considered from the same underlying protein across QC spectra. From this point of view, we have formulated the peak matching problem as a maximal clique finding problem. In the general case, this problem is known to be NP-hard. However, we are able to exploit the special structure of our data to devise a clique finding algorithm that is solved in polynomial time. For details, the interested reader is referred to the code accompanying our manuscript.

4.3 Results

The quadratic variance function can be reliably estimated from the gaps of the spectra in between the peaks. This is shown by the graph in Figure 14. This confirms that the area interspersed between peaks in SELDI follow the quadratic variance model.

LibSELDI consistently finds more peaks that Ciphergen Express does. This is shown in Figure 15. LibSELDI finds between 100 to 200 peaks per spectrum in the 3 to 30 kDa

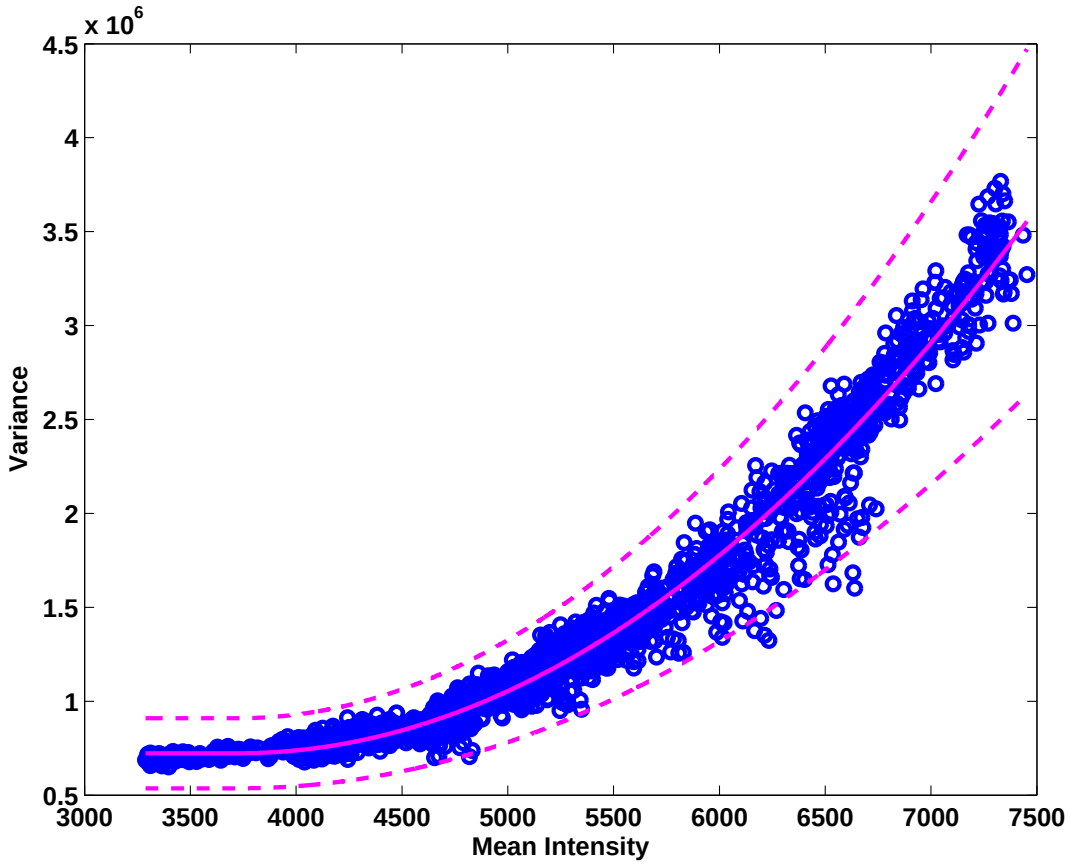

Figure 14: Variance of measurements are a quadratic function of the mean. The blue circles indicate mean/variance points estimated from regions in between peaks in the spectra. The solid magenta line is the best fit quadratic variance function, while the dotted magenta lines indicate plus/minus one standard error.

range. This is a reasonable number to expect for such a range of mass values. Note that the number of peaks predicted is always less than the peak capacity as defined by [18].

LibSELDI finds more than four times as many reproducible peaks as Ciphergen express. This is illustrated in Figure 16. Most biologists are interested in the number of peaks found with a prevalence greater than 80%. LibSELDI finds 84, while Ciphergen express only finds 18.

Mean peak heights and peak height variances are consistent with the quadratic variance model in the majority of cases. We restrict our analysis to peaks appearing in at least 50% of the spectra (guaranteeing at least n=16 for sample means and variances). We illustrate this for the range of intensity values encompassing most of our peaks in Figure 17. For the

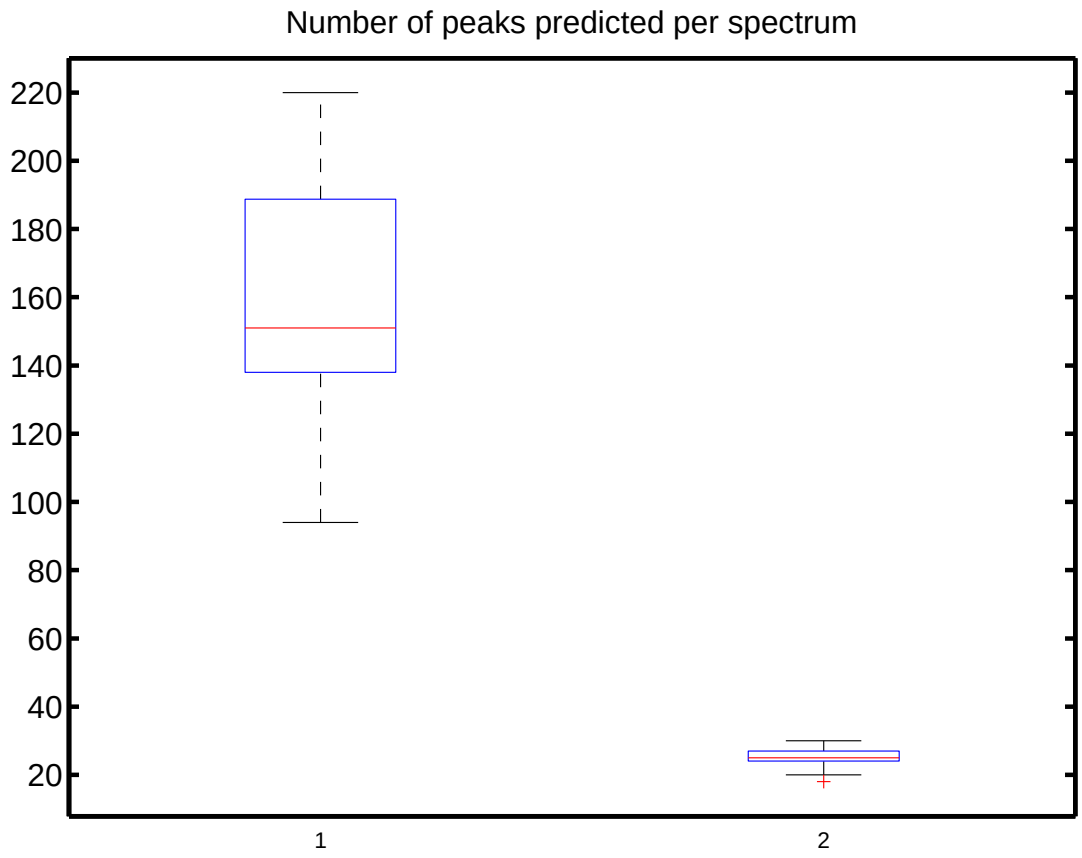

Figure 15: LibSELDI finds more peaks per spectrum than Ciphergen Express. Box-plots are shown with the y-axis indicating number of peaks predicted in a QC spectrum. The predictions corresponding to LibSELDI is indicated by a 1, while Ciphergen is demarcated by a 2 on the x-axis.

few cases with peaks of very high mean intensity occurring in the spectra, the model does not fit. This is shown in Figure 18.

The CV values of peak heights observed are consistent with the predictions of the quadratic variance model in most cases. We illustrate this in Figure 19. Similar to what we saw in Figure 18, the model breaks down for peaks at very high mean intensity, which are a small minority of our observations. This breakdown at high intensities is shown in Figure 20. Note, however, that the predictions are still bounded below the large μ CV approximation predicted by the model in Eq. (27).

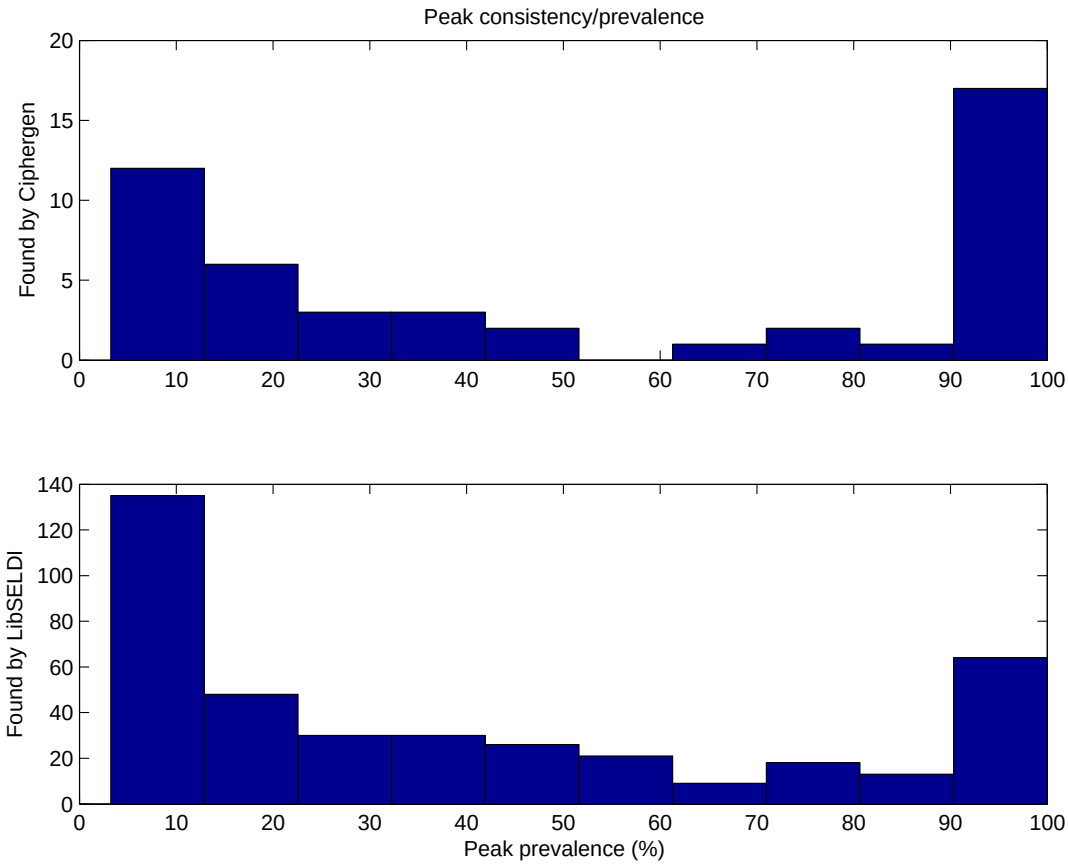

Figure 16: LibSELDI finds more reproducible peaks than Ciphergen Express. LibSELDI finds 84 peaks occurring in at least 80% of our QC spectra, while Ciphergen finds only 18 such peaks.

4.4 Discussion

The quadratic variance model of measurement for SELDI explains most of our observations about the reproducibility of peaks in this study. Examining Figures 14, 17, 18, 19, and 20, a picture starts to emerge of a measurement model that is constant variance for mean intensities below 3700, quadratic variance between 3700 and 12000, and possibly transitioning to constant variance for very high intensities above 12000. A large majority of the peak heights from the pooled mucous QC samples were observed in the quadratic variance region. More work is needed to understand the behavior at very high intensities. This model may also be generalizable to MALDI since the technology is actually very similar SELDI.

The LibSELDI preprocessing approach, taking into account the quadratic variance of

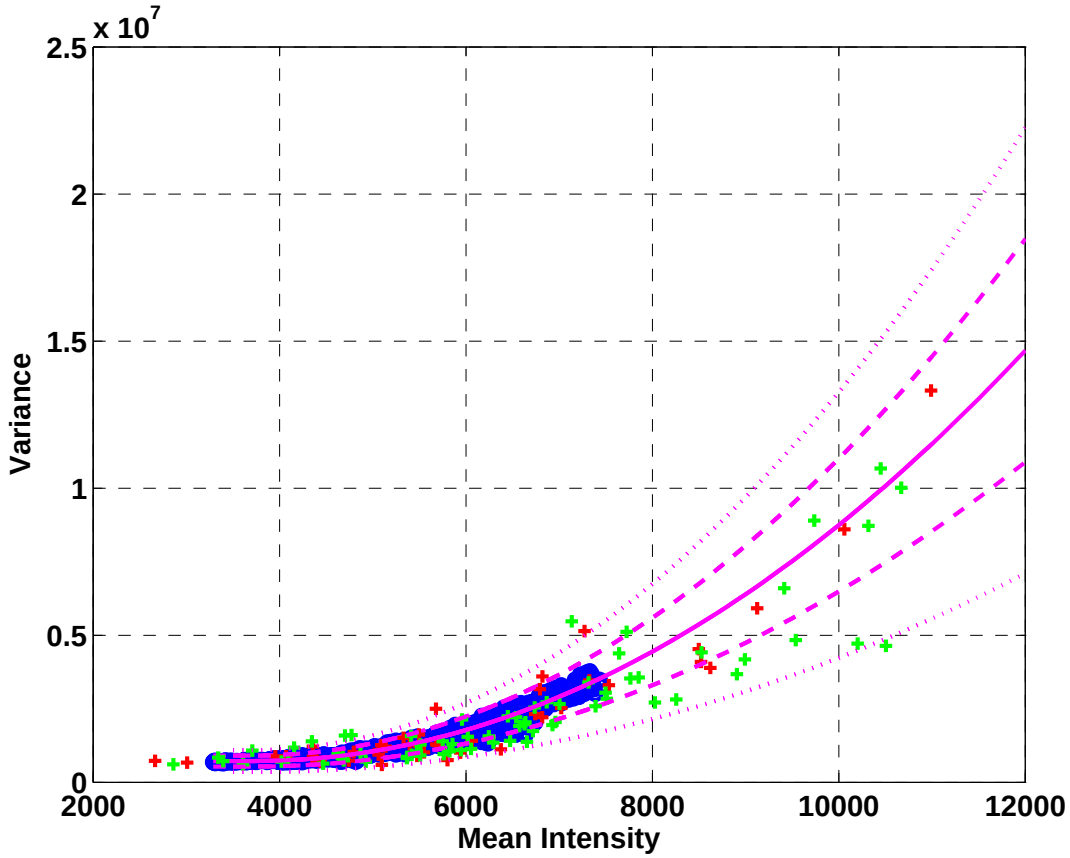

Figure 17: Mean peak heights and peak height variances are consistent with the quadratic variance model for most peaks. The blue points indicated the mean/variance pairs from non-peak regions used to estimate the model. The red plus symbols corresponding to peaks occurring in at least 80% of QC spectra, while the grean plus symbols indicate peaks occurring in 50% - 80% of QC spectra. The dashed (dotted) magenta lines indicates one (two) standard errors from the mean, respectively.

the data, demonstrates certain strengths. By measuring the characteristics of the noise associated with a set of experimental conditions and machine settings, LibSELDI is able to adapt to changing noise/background characteristics. The peak matching algorithm formulated as a clique-finding algorithm performs favorably. The sensitivity demonstrated for finding peaks occurring in more the 80% of the spectra is impressive– finding more than four times as many as Ciphergen (84 peaks versus 18). Further, the protein estimates/peaks found by the model had mean peak heights, variances, and CV's that are consistent with what would be predicted by the model. Thus the quadratic variance function estimation, estimated rather simply, says something about how reproducible our peaks are going to be

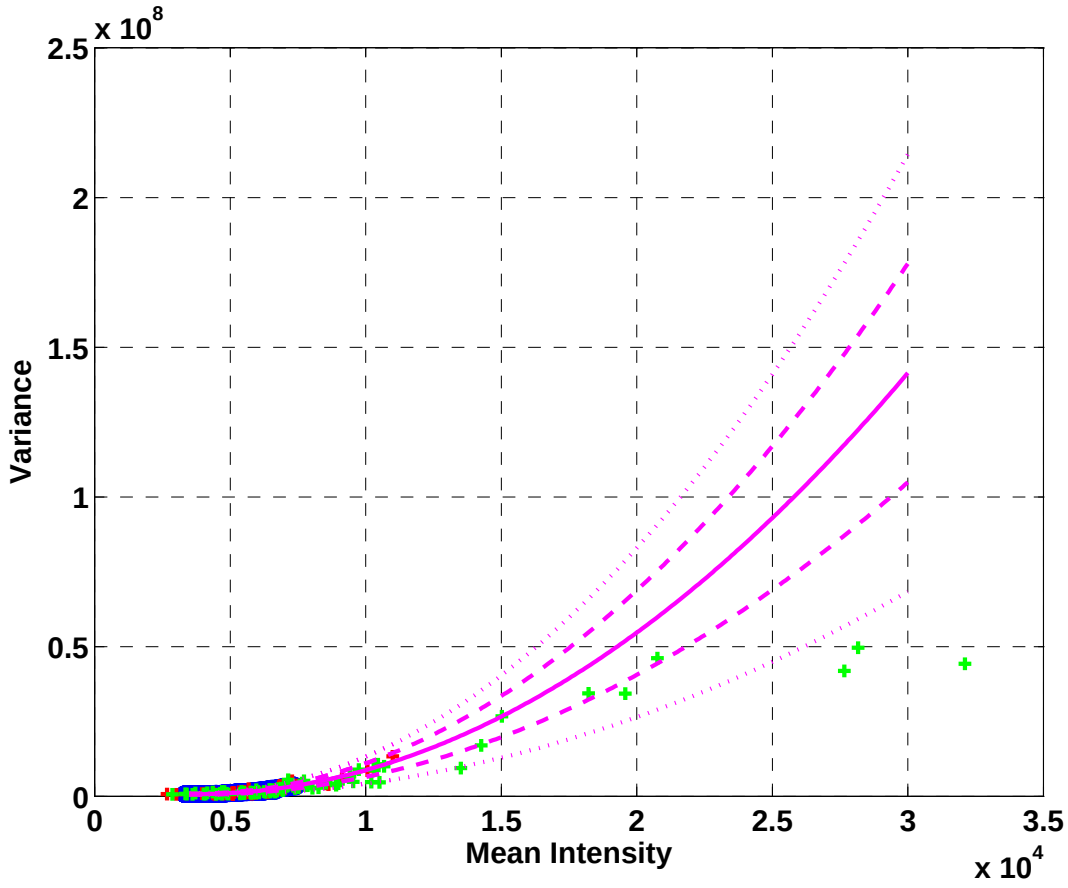

Figure 18: Mean peak heights and peak height variances for very large mean height values are not consistent with the quadratic variance model. The blue points indicated the mean/variance pairs from non-peak regions used to estimate the model. The red plus symbols corresponding to peaks occurring in at least 80% of QC spectra, while the green plus symbols indicate peaks occurring in 50% - 80% of QC spectra. The dashed (dotted) magenta lines indicates one (two) standard errors from the mean, respectively.

in advance of an experiment run.

The are limitations to the LibSELDI approach, however. The modified Antoniadis-Sapatinas algorithm at the core of LibSELDI is computationally intensive– requiring approximately 7 minutes of processing time per spectrum on a Dell Precision 690 with 12 Gigabytes of RAM. The model is likely to break down at m/z close to around 2.5kDa where the baseline saturates due to non-linearities introduced by the detector saturating. In this region, Figure 17 implies that the algorithm may over-estimate the noise thus possibly over-smoothing and missing a peak.

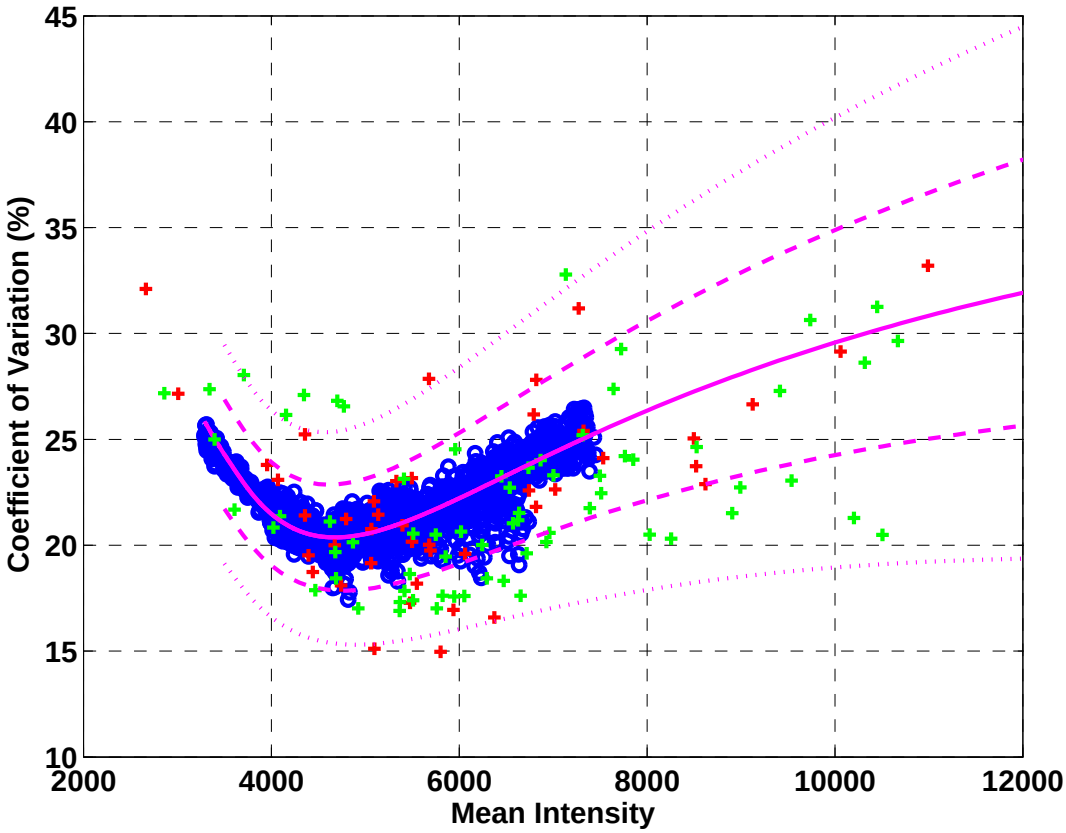

Figure 19: Observed CV values of peaks are consistent with the quadratic variance model in most cases. The blue points indicated the mean/CV pairs from non-peak regions used to estimate the model. The red plus symbols corresponding to peaks occurring in at least 80% of QC spectra, while the grean plus symbols indicate peaks occurring in 50% - 80% of QC spectra. The dashed (dotted) magenta lines indicates one (two) standard errors from the mean, respectively.

Although LibSELDI demonstrated superior peak detection performance on our cervical mucous QC data, Ciphergen Express still has certain strengths. First and foremost, it is a mature software package easily usable by the community. In contrast, LibSELDI is still in development and thus some MATLAB programming expertise is required to use this package in addition to the availability of good computing resources. However, the main drawback of Ciphergen in this paper is its lack of sensitivity as illustrated clearly in Figures 15 and 16.

The quadratic variance model has been demonstrated to explain the variation observed in cervical mucous QC data, leading to more reproducible peak detection and predicting

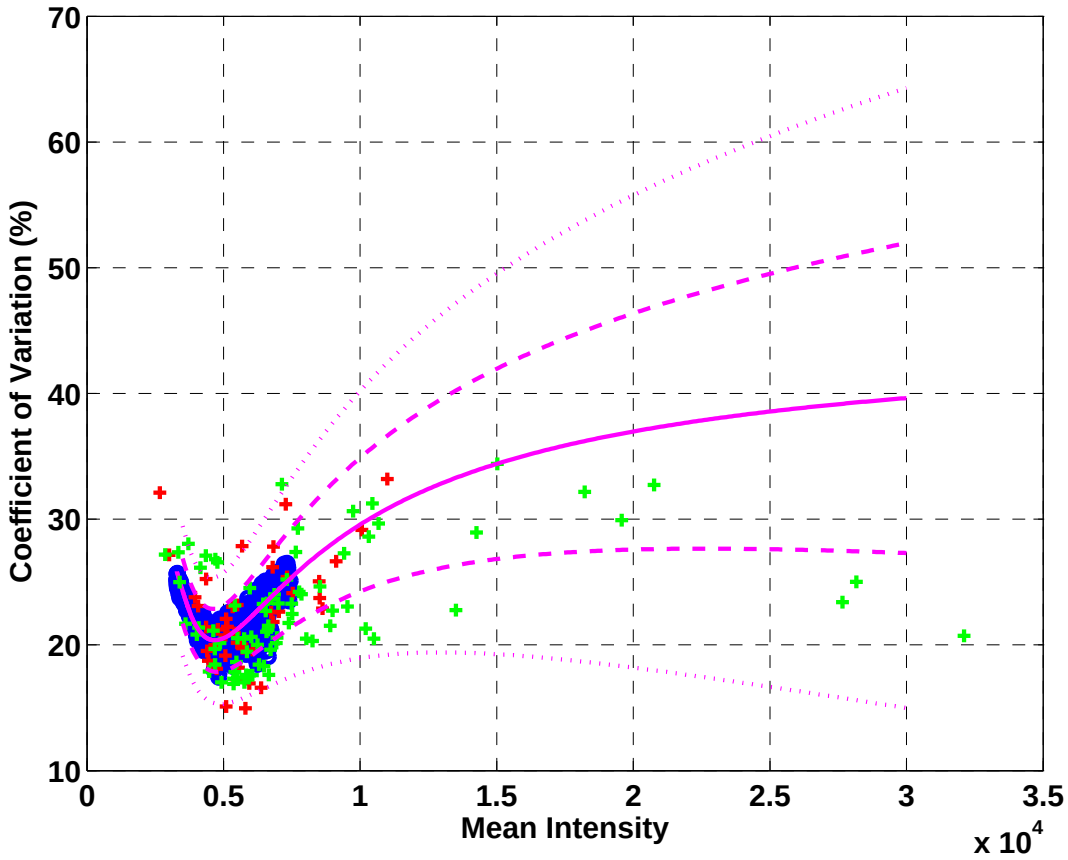

Figure 20: Observed peak height CV values for peaks at very high intensity are not consistent with the quadratic variance model. The blue points indicate the mean/CV pairs from non-peak regions used to estimate the model. The red plus symbols corresponding to peaks occurring in at least 80% of QC spectra, while the grean plus symbols indicate peaks occurring in 50% - 80% of QC spectra. The dashed (dotted) magenta line indicates one (two) standard errors from the mean, respectively.

the CV of the peak heights successfully.

CHAPTER V

CONCLUSIONS

5.1 Contributions

If there is a single item of knowledge that is to be remembered from this dissertation it is this: the acquisition of SELDI mass spectrometry measurements follow the quadratic variance model. From this, almost everything else can be deduced including the preprocessing techniques used and our expectations about the reproducibility of peaks in the data.

The full list of contributions of this work are:

- We produced one of the first large-scale comparisons of SELDI preprocessing algorithms to date consisting of the comparison of 9 different approaches on a test set of 10,000 spectra. Several benchmarking metrics were used for this study taking into consideration the needs of both computational scientists and laboratory scientists (which are in fact often quite contradictory). These simulations consumed more than a year of CPU time. This resulted in a publication in the journal Proteomics [39], with a figure from our publication being chosen as one of the figures on the cover of the issue.

- We proposed a natural exponential family with quadratic variance function model for the acquisition of SELDI data based on an analysis of buffer/matrix only spectra. We conjecture that this model captures the behavior of the detector.

- We introduced the modified Antoniadis-Sapatinas wavelet denoising algorithm that takes into account the quadratic variance model to denoise each spectrum by the right amount [40].

- We have developed a collection of preprocessing algorithms, including the modified Antoniadis-Sapatinas algorithm, implemented in Matlab in a package we call LibSELDI. We have contributed 30,000+ lines of code to LibSELDI (printed out, this

would span 500+ pages). LibSELDI is licensed under v3 of the GNU public license and is intended to form the core of a free and open sourced replacement for the expensive Ciphergen Express software sold with SELDI.

- We have adhered to the concept of reproducible computational research in the sense that all of our publications are released with the code and data necessary to reproduce our figures by simply running a script named "pulishedfigures.m".

- Lastly, we have shown that the quadratic variance model, based on measuring blank regions of a spectrum, explains the observed peak behavior of pooled cervical mucus QC spectra from data being used in an early cancer detection project here at the CDC. For these spectra, LibSELDI finds 84 reproducible peaks in comparison to 18 for Ciphergen Express.

The notion of quadratic variance of the measurements is fundamental and has the potential to impact numerous publications. In Figure 21, we show the trends in publications for SELDI in the past 10 years. Nearly all of these papers preprocess SELDI mass spectrometry data and thus may be impacted by our work, provided that further validation of the model continues to vindicate its use.

5.2 Publications

The work related to this thesis has resulted in several publications, which are listed below.

5.2.1 Journals

[J1] **V. A. Emanuele II** and B. M. Gurbaxani, "Benchmarking Currently Available SELDI-TOF MS Preprocessing Techniques," *Proteomics* 2009 Apr; 9 (7), pp. 1754-62. **Figure from paper selected for cover of issue.**

[J2] **V. A. Emanuele II** and B. M. Gurbaxani, "Quadratic variance models for adaptively preprocessing SELDI-TOF mass spectrometry data" accepted subject to revisions in *BMC Bioinformatics*

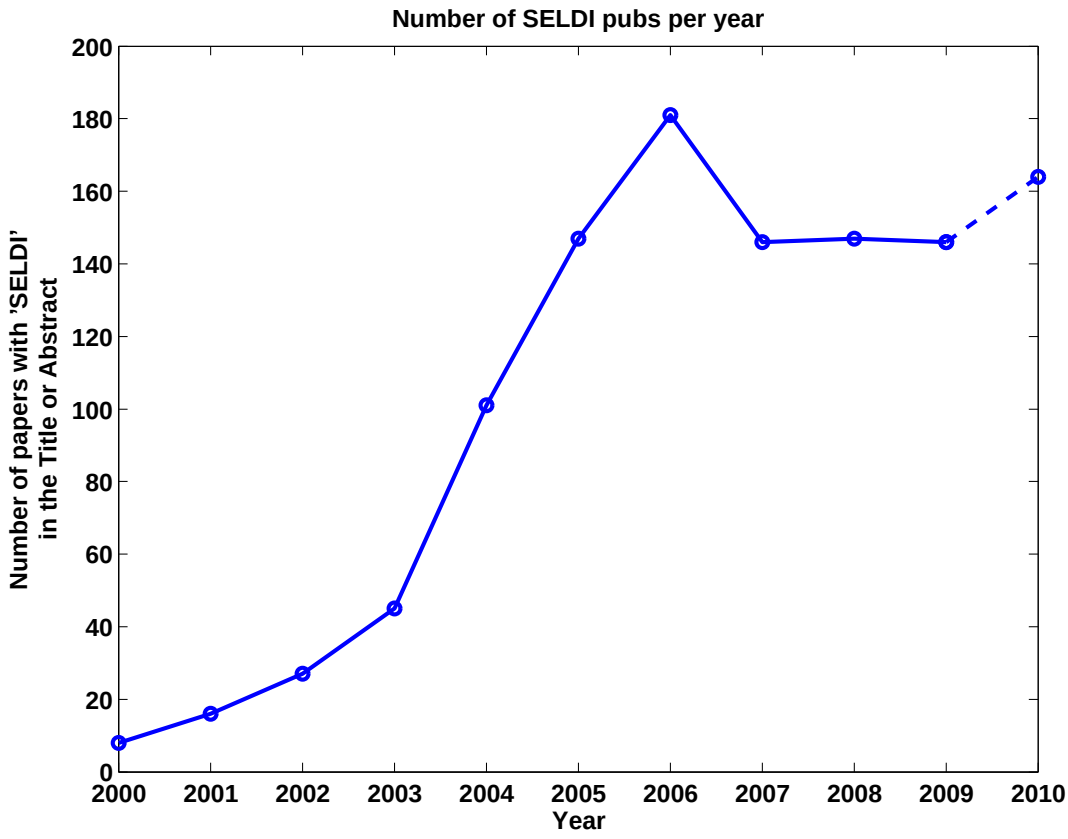

Figure 21: Publication trends for SELDI for the past 10 years. This is a figure that was generated from inspiration from the analogous figure in [118].

[J3] **V. A. Emanuele II**, G. Panicker, B. M. Gurbaxani, and E. R. Unger, "Explaining reproducibility of peaks in SELDI mass spectrometry: the quadratic variance model" in preparation with co-authors for *Clinical Chemistry*. Expected submission date: September 2010.

5.2.2 Conferences

[C1] **V. A. Emanuele II** and B. M. Gurbaxani, "Modeling uncertainty in SELDI mass spectrometry with applications to biomarker discovery in complex biological media", poster presented at the *National Center for Enteric, Zoonotic, and Infectious Diseases Science Summit*, Centers for Disease Control and Prevention, Atlanta, GA, August 24, 2010.

[C2] G. Panicker, **V. Emanuele II**, B. Gurbaxani, D. Lee, and E. Unger., "Enroute to

protein biomarker discovery for early-detection of cervical cancer,", Poster presented in *CDC Celebrates 10 Years of Public Health Genomics: Translating Gene Discoveries into Population Health Benefits*, Atlanta, Ga, January 23, 2008.

[C3] **V. A. Emanuele II** and B. M. Gurbaxani, "Comprehensive Performance Analysis of SELDI-TOF Mass Spectrometry Signal Processing Methods," Poster in *2nd Annual Computational and Systems Biology Symposium*, University of Georgia, March 23rd, 2007.

[C4] **V. A. Emanuele II** and B. M. Gurbaxani, "SELDI-ToF based protein profiling for early detection of cancer and biomarker discovery: an evaluation of currently available signal processing platforms," Poster in *AACR Advances in Proteomics in Cancer Research*, Amelia Island, FL, February 27-March 2, 2007.

[C5] **V. A. Emanuele II**, V. Olman, B. Yan, Y. Xu, and G. T. Zhou, "An Approximate Bayesian Detection Scheme with Applications to Tandem Mass Spectrometry Data Analysis," in *Proceedings of the 12th Digital Signal Processing Workshop (DSP 2006)*, pp. 550-560, Jackson Hole, WY, September 2006.

[C6] B. Yan, G. T. Zhou, P. Wang, Z. Liu, **V. A. Emanuele II**, V. Olman, and Y. Xu, "A Point-Process Model for Rapid Identification of Post-Translational Modifications," in *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pp. 327-338, Wailea, Maui, Hawaii, January 3-7, 2006.

[C7] **V.A. Emanuele II**, T. T. Tran, and G. T. Zhou, "A Fourier Product Method for Detecting Approximate Tandem Repeats in DNA," in Proceedings of the *IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, July 17-20, 2005.

[C8] T. T. Tran, **V. A. Emanuele II**, and G. T. Zhou, "Techniques for Detecting Approximate Tandem Repeats in DNA," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 17-21, 2004

APPENDIX A

ADDITIONAL COMMENTS AND NOTES SUPPLEMENTING THE BENCHMARKING STUDY

We have approximately 2 GB of Matlab code, perl scripts, and data that we will make available upon request via FTP. This code by itself could be used as a toolbox of sorts for taking results, formatted properly, from a new program to be benchmarked and performing the same analysis we have published. We include a few scripts that will regenerate all the figures and tables presented in the published paper and supplementary information herein. There is additional information included in this larger supplement, including tables of parameter combinations used and the corresponding operating points observed, that may be of interest to the community.

A.1 Choosing algorithms in the study

Our primary goal was to benchmark algorithms and techniques that are accessible to practicing scienctist. With this in mind, we decided that the SELDI preprocessing platforms must satisfy the following criteria to be included in the study:

1. **Availability:** The software must be available. Available means that we must be able to download the source code and install the package ourselves at no charge. This *does not* include software that authors are not willing to share with us.

2. **Output Format:** The software must produce a list of m/z values detected as protein mass values detected in the group of spectra, along with corresponding intensity estimates.

3. **Usability:** The software program must be usable by someone who is *not* an expert in computer science and/or bioinformatics.

A.2 Model for MALDI/SELDI Protein Profiling Data

We briefly describe the model used in the MS simulation engine, emphasizing the aspects of the model that our critical for our work (see [18, 82] and the accompanying supplementary information for the original description given by Morris, Coombes, and colleagues).

A.2.1 Sample Collection

The MS simulation engine uses a simple model for sample collection that, in our opinion, is an accurate reflection of the first order characteristics observed in our SELDI data. A protein population in this model is completely characterized by modeling the distribution of four quantities, $p, log(x), a, s$. The parameter p is the protein prevalence [18], which is the probability that a protein occurs in a spectrum drawn from this population. This models the common observation that peaks at a corresponding m/z often occur in some fraction of the spectra in one's data, but often do not occur in all spectra. This distribution of protein prevalences in our population is modeled using the beta distribution [82].

The vector parameters $\boldsymbol{\theta} = (\log(x), a, s)$ are the logarithm of the protein mass, the mean log-intensity value of the peaks in the spectra generated by the corresponding protein (called abundance), and the standard deviation of the log-intensity of the peaks in the spectra generated by the corresponding protein. Morris and colleagues have assessed that a multivariate normal distribution is sufficient for accurately describing $\boldsymbol{\theta}$ [82].

The parameters $\{p, \boldsymbol{\theta}\}$ in this model and their corresponding distribution functions represent a characterization of the behavior of all the proteins in this population as a whole. For example, we could imagine different parameterizations $\{p, \boldsymbol{\theta}\}$ for serum samples from prostate cancer patients and healthy men. Thus we are positing that each time we ask 100 prostate cancer patients to come to the clinic, it amounts to observing a sampling from p and $\boldsymbol{\theta}$ for each of, say, 150 proteins. Thus, for this sample population of 100 prostate cancer patients, we have values $p_1, \ldots, p_{150}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{150}$ for the proteins that could be observed in this group. For each patient, we evaluate the possibility of each of these 150 proteins being observed by sampling from a uniform random variable U on $[0, 1]$, where the protein is observed if $u_j < p_j$. For each protein that is observed in this spectrum, we next sample

from the corresponding normal distribution with mean a_j and standard deviation s_j that describes the behavior of the log-intensity, and map this value to an ion count (number of proteins present at this corresponding mass for this patient) as is described in [82] and its supplementary information. This model appears to capture the first order variations in data collected from patients who represent the same population (e.g. prostate cancer patients), sampled from different clinics.

A.2.2 Ionization/Desorption

The largest factor affecting variability in flight time for the same protein is arguably its initial velocity upon ionization. Interestingly enough, the initial velocity off of the sample plate is roughly independent of mass and is modeled with a normal distribution [64].

A second order effect is the variation of mass for a given protein due to the distribution of isotopes of common elements such as carbon, nitrogen, and oxygen. This variation is modeled in the simulation engine using a simple binomial model [18].

A.2.3 Analysis and Detection

Once the mass and initial velocity are sampled for all the particles corresponding to protein A, the time of flight is calculated deterministically, assuming typical machine settings and geometry for a low resolution MALDI [18]. When the virtual spectrum is generated, an exponentially decaying baseline signal and white Gaussian noise process are added to the spectrum to account for detector saturation and electronic noise, respectively. The i.i.d. Guassian noise process is zero mean with a standard deviation of 66. For details of what the parameter settings are for the simulated data, see [18, 82] and the provided data.

A.3 Notes on how parameters were chosen for each algorithm

In selecting the parameter ranges to explore for each program, we adhered to a few basic rules.

1. When code was available demonstrating how to use the programs, we attempted to mimic the example in our implementation.

2. If certain ranges of values were recommended in published literature or documentation accompanying the programs, we followed these guidelines.

3. We explored as many parameter combinations as possible given the runtime and efficiency of the algorithm.

The structre of some programs, such as MassSpecWavelet [35] and Mean Spectrum [82], lent themselves to easy evaluation. For these programs, we could simply order all the protein predictions by their SNR value and move the threshold from highest to lowest to generate a huge number of operating points quickly. These sort of tricks have been discussed in [43].

Other programs, such as Bioconductor PROcess [49] and Genepattern, were extremely computationally intensive and more challenging to evaluate. To evaluate these slower programs we used a small computing cluster containing 8 nodes.

Unfortunately, Ciphergen Express [46] has no scripting capability and the evaluation of this program was laborious. In order to decide which parameter combinations to try for this program, we selected five parameter combinations that gave us rougly 20, 50, 150, 300, and 600 protein predictions for Dataset 1. This way, although we generated only 5 operating points, it was spread out over a range of operating capabilities.

The tables of parameter values and their corresponding operating points for each program on each dataset is contained in the supplementary datasets available via FTP (see README.txt for exact location).

A.4 Example Predictions

Finally, we present some example predictions of the top performing programs to facilitate a discussion of strengths and weaknesses of the algorithms. We focus our analysis in this section on three preprocessing programs: Ciphergen Express, Mean Spectrum, and MassSpecWavelet. Ciphergen Express and Mean Spectrum were the top two programs with respect to the PAUC measure, while MassSpecWavelet was the top finisher with respect to MEANTPR. If one were to consider the sum of the ranks of the programs with respect to PAUC and MEANTPR as a measure of a program's potential, these programs would be the top three.

Figure 22 shows example peak predictions for Dataset 10 of the simulated data. Each row

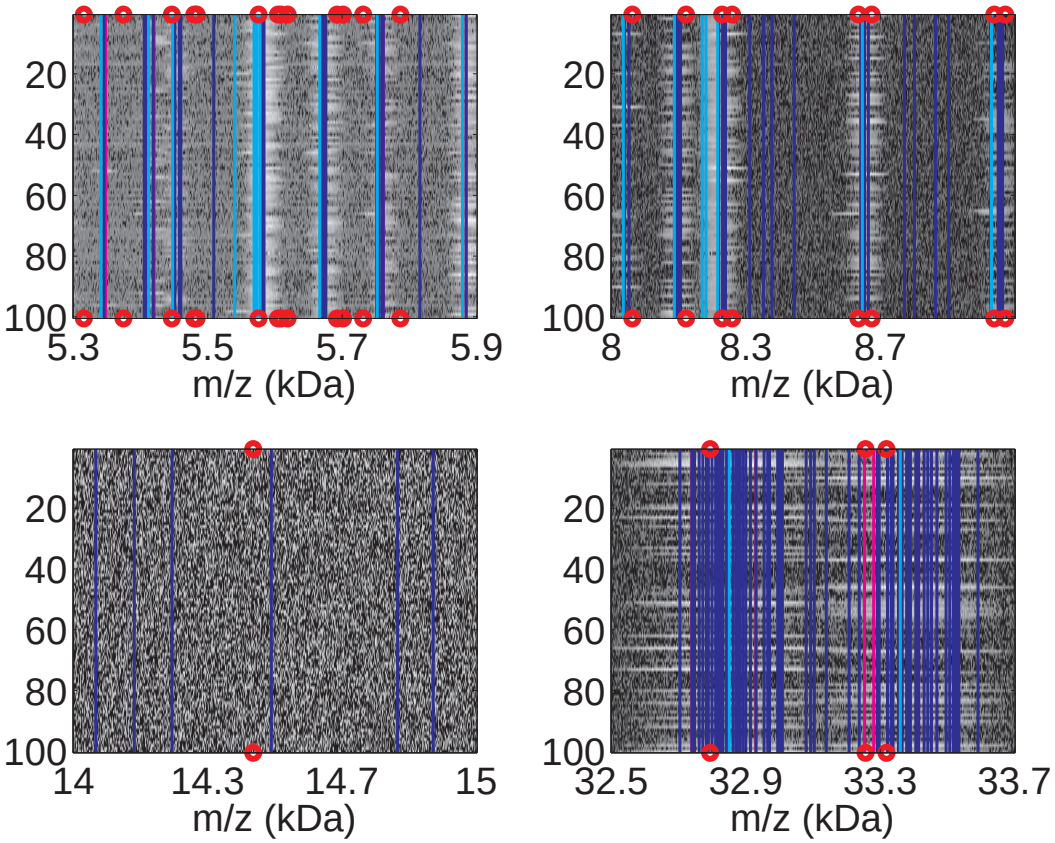

Figure 22: Example peak predictions of the top three programs on Dataset 10. Ciphergen Express, MassSpecWavelet, and Mean Spectrum are shown in purple, light blue, and dark blue, respectively. Note that all 100 spectra in Dataset 10 are displayed here as a heat map. The red dots indicate location of the actual protein m/z value used in the simulation.

corresponds to a spectrum from the dataset, with the log-intensity of the spectrum displayed as a grayscale heatmap. In the figure, white represents high intensities and black represents zero intensity. The red circles on the x axes represent the m/z values of known virtual proteins in the dataset, and the vertical lines represent protein predictions by the various programs. Predictions made by Ciphergen Express, Mean Spectrum, and MassSpecWavelet are shown in purple, dark blue, and light blue respectively.

Briefly observing all of the figures, we can make a couple general observations. First, it is clear that Ciphergen Express seems to only predict the existence of a protein when it is very sure. Unfortunately, the Ciphergen Express preprocessing algorithms are largely

a "black box" of sorts. Thus, it is difficult to say why the program performed as it did since the details of algorithms are not completely available. The conservative nature of the Ciphergen Express program is evident in its top ranking with respect to the PAUC metric. We regard this as a virtue.

On the other hand, Mean Spectrum and MassSpecWavelet do well at recovering the protein m/z values from the data at the expense of a multiplicity of predictions. Indeed, Figures 22a, 22b, and 22d illustrate that these programs tend to predict clusters of peaks in the area of a true protein. This is perhaps a consequence of the way both programs apply the wavelet transform. Since lower mass peaks tend to have higher intensities, the corresponding wavelet coefficients for the wavelet bases at fine resolution turn out to be large, thus making these two wavelet-based techniques more sensitive than expected to narrow spurious noise peaks. One possible way to improve the wavelet based methods could be to incorporate knowledge of the expected peak resolution for SELDI in an adaptive, m/z dependent, wavelet coefficient thresholding.

We have provided matlab code used to generate the figures used in this publication and additional information in the supplementary information.

A.5 Information about Matlab code, perl scripts, and data supplement available via FTP

Note below is a reprint of the README.txt file included with the 2GB supplement.

```
------------------------------------------------------------
S.1 - How to reproduce the figures published in this paper
------------------------------------------------------------
Requires: Matlab v.7.3 (R2006b) or later and read/write access to the
MATLABROOT directory.

1) These directions assume you have followed the accompanying directions with
   this zip file and successfuly unzipped this file in the 'work' subdirectory
   of the MATLABROOT folder. This then creates the EmanueleSuppl subdirectory
```

in the work subdir.

2) Start matlab.

3) Set current working directory to the MATLABROOT/work/EmanueleSuppl dir (on linux) or MATLABROOT\work\EmanueleSuppl on a Windows machine. This can be done as follows.

Ex:

```
>> cd( fullfile( matlabroot, 'work', 'EmanueleSuppl' ) )
>>
```

4) At the Matlab command prompt, enter 'publishedfigures'. This generates the from our publication.

Ex:

```
>> publishedfigures
publishedfigures.m: Note: Figure 1 was hand-made (nothing to simulate)
publishedfigures.m: Generating figure 2...
publishedfigures.m: Done!
   (and so forth as the program continues to run)
```

5) Next, type 'publisheddatafast' at the Matlab command prompt. This displays the stats used to generate the tables from the publication.

6) Using a similar procedure, the supplementary info/figures can be generated as well. See Table of Contents listing in the next section.

S.2 - Table of Contents w/ descriptions

We briefly describe the organization of the supplementary info in our paper. Note that we do not list every directory/subdirectory contained here but only highlight what we believe to be the main ones of interest

Scripts used to reproduce results

publisheddatafast.m - Displays info used to generate tables in the publication.

publishedfigures.m - Runs code that produces the figures used in the publication.

****Other scripts****

WARNING: For the complete supplementary info necessary to run these scripts, it is necessary to email the authors to get additional data (>2GB worth of data).

supplfigures.m - Runs code to generate supplementary info and figures (may take a while)

publisheddata.m - Same as publisheddatafast.m, except the analysis is actually performed rather than loading a file containing the results of the analysis done previously. WARNING: This may take a very, very, long time depending on your system.

publishedfiguresprep.m - Runs simulation steps to produce 'matfiles/figuredata.mat', a file used by publishedfigures.m. Not particularly necessary to do this unless you were wondering where this file ('matfiles/figuredata.mat') came from.

Directory Contents

algorithmparams/ - Contains tables of parameters explored for each algorithm, as well as the corresponding operating point. Consists of one subdirectory for each algorithm, with each subdirectory containing one table (csv file) for each dataset.

data/simulated/ - contains the simulated datasets tested, as downloaded from http://bioinformatics.mdanderson.org/Supplements/Datasets/Simulations/index.html

discussion/ - Contains written documents supplementing the main content of the paper.

figures/published/ - where the results of 'publishedfigures.m' end up... the reproduced figures from our publication

figures/supplement/ - location of additional figures to compliment those in the publication.

lib/ - location of code used in benchmarking simulations.

lib/matlab/classes/ - the heart of the code used in benchmarking simulions (using object-oriented matlab).

lib/matlab/mytools/ - Add'l functions developed for use in benchmarking simulations (typically not object-oriented).

lib/matlab/scripts/ - Major scripts used to generate results in paper, illustrating the use of the classes contained in lib/matlab/classes.

```
lib/matlab/toolbox/ - location of 3rd party programs used in the benchmarking
simulations.  Most of these were downloaded from Matlab Central File Exchange.

matfiles/ - location of .mat files containing simulation info used to construct
publication figures.

--------------------
End Table of Contents
--------------------
```

A.6 Operating Characteristics

Here we show the mean operating characteristics with standard error bars for each of the programs tested.

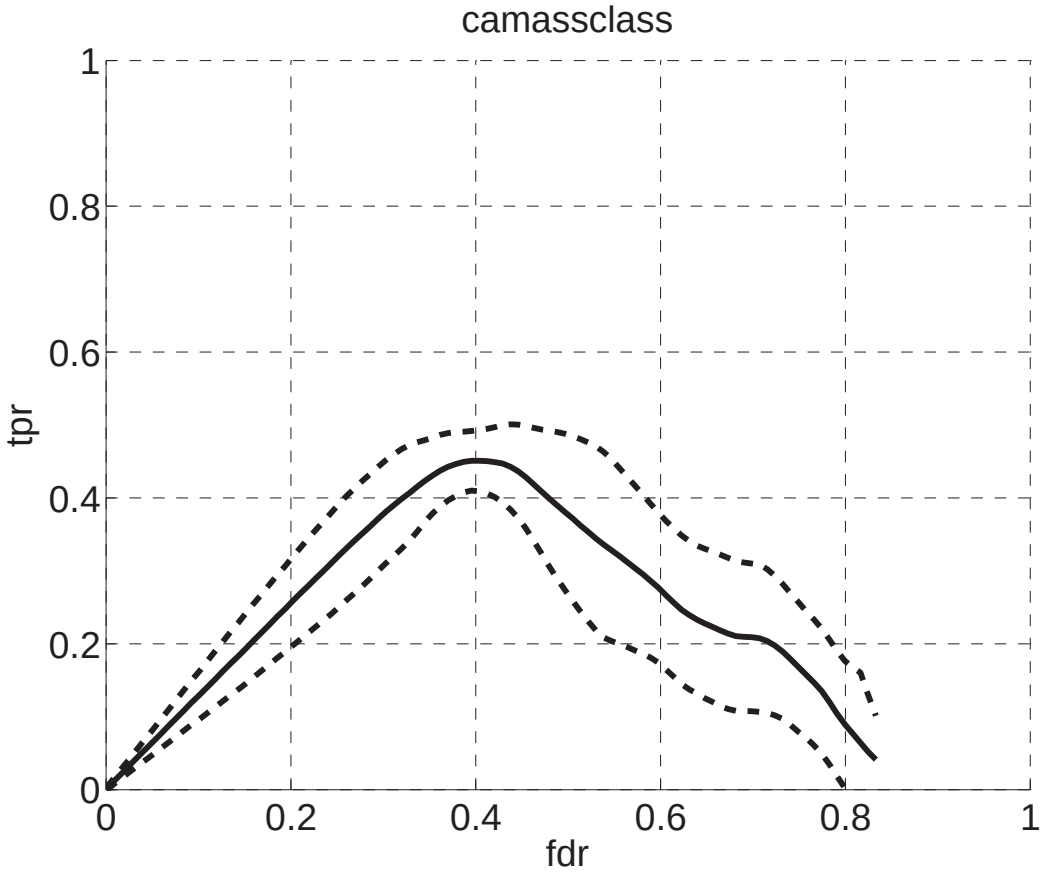

Figure 23: Operating characteristic: caMassClass

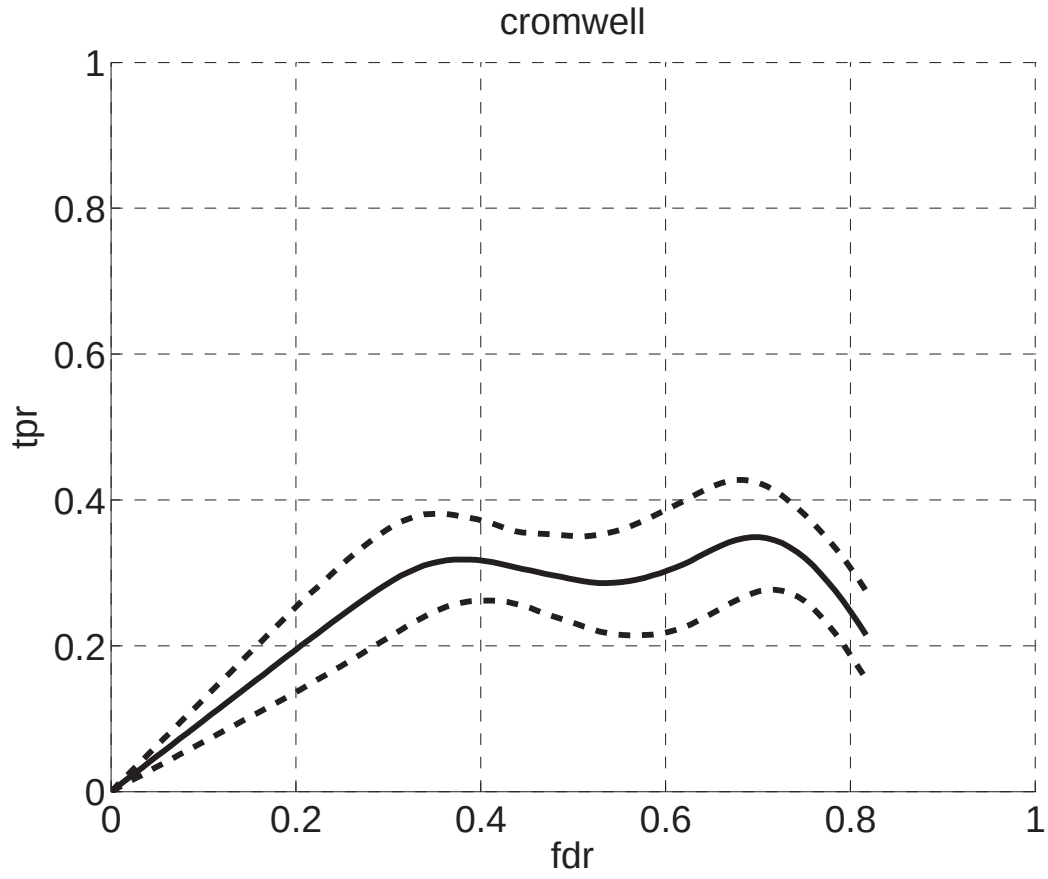

Figure 24: Operating characteristic: Cromwell

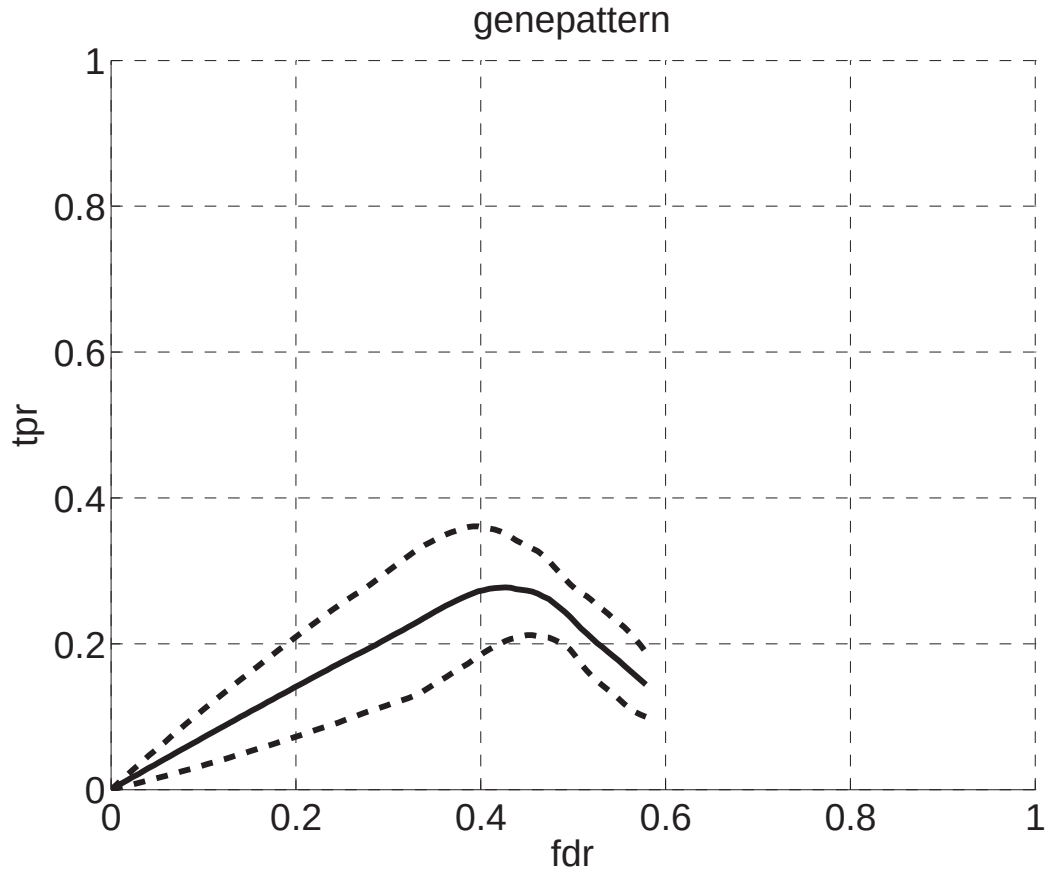

Figure 25: Operating characteristic: GenePattern

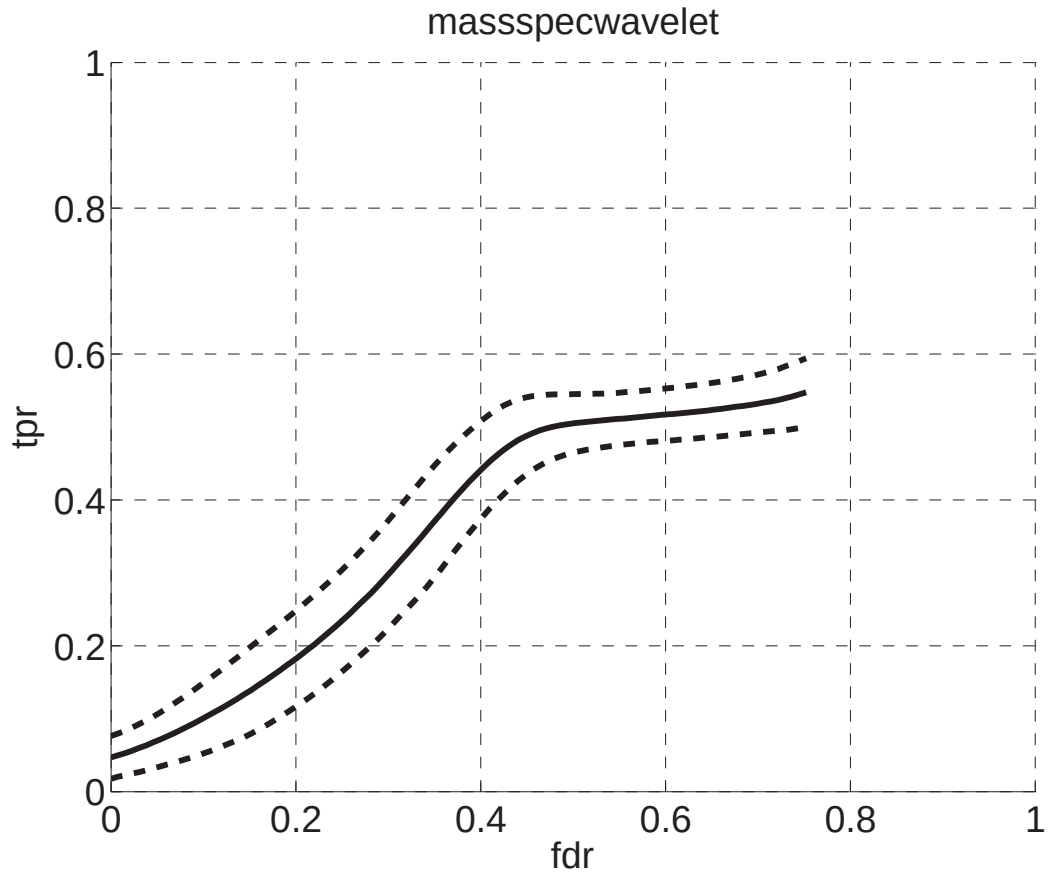

Figure 26: Operating characteristic: MassSpecWavelet

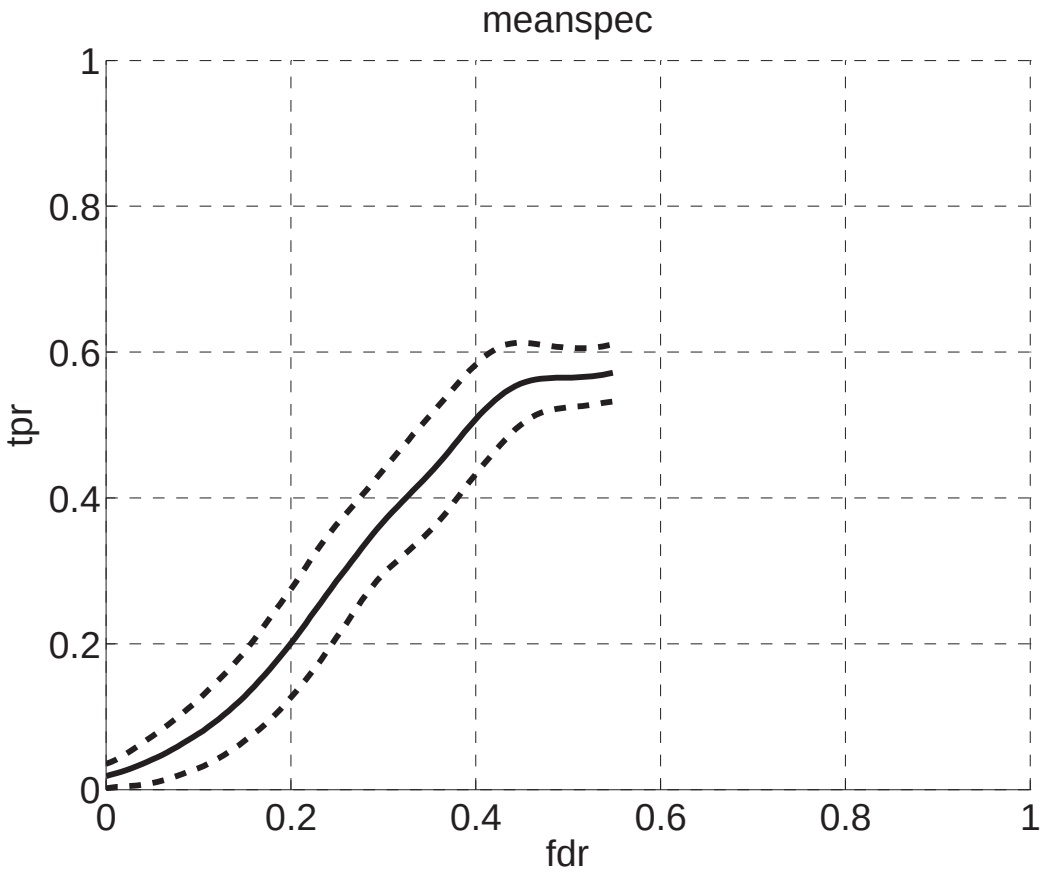

Figure 27: Operating characteristic: MeanSpectrum

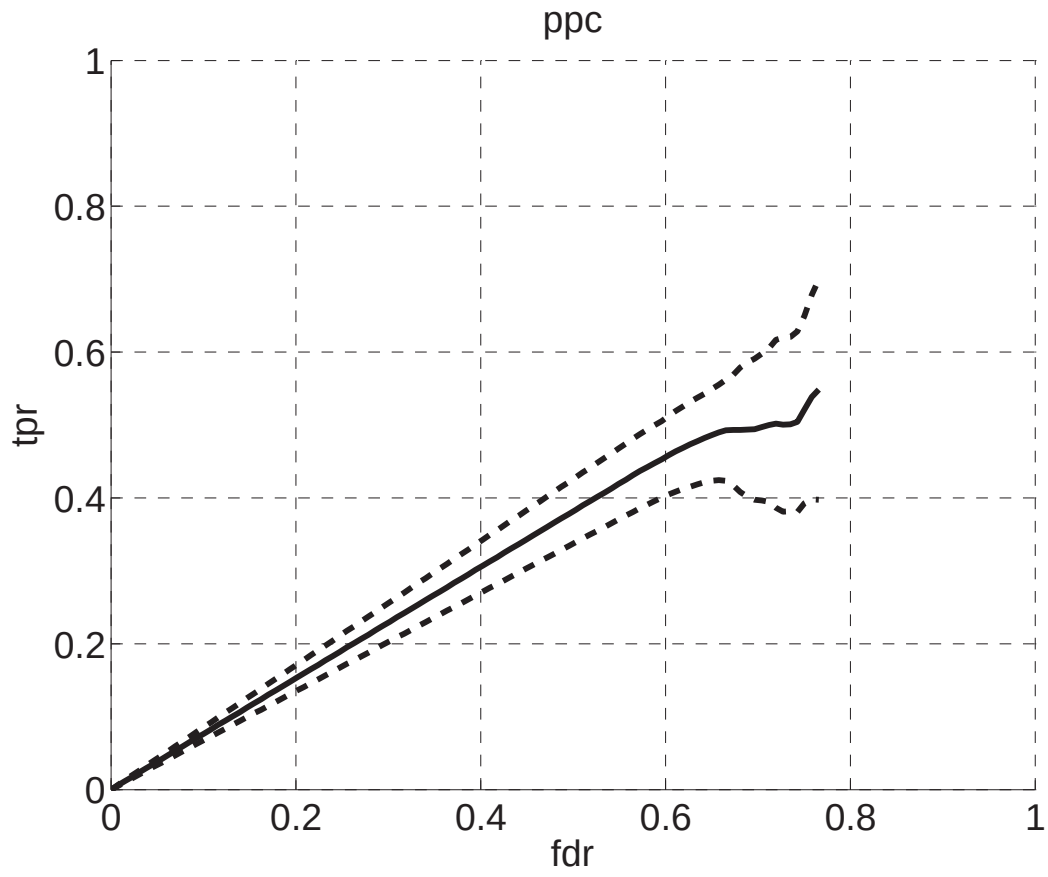

Figure 28: Operating characteristic: PPC

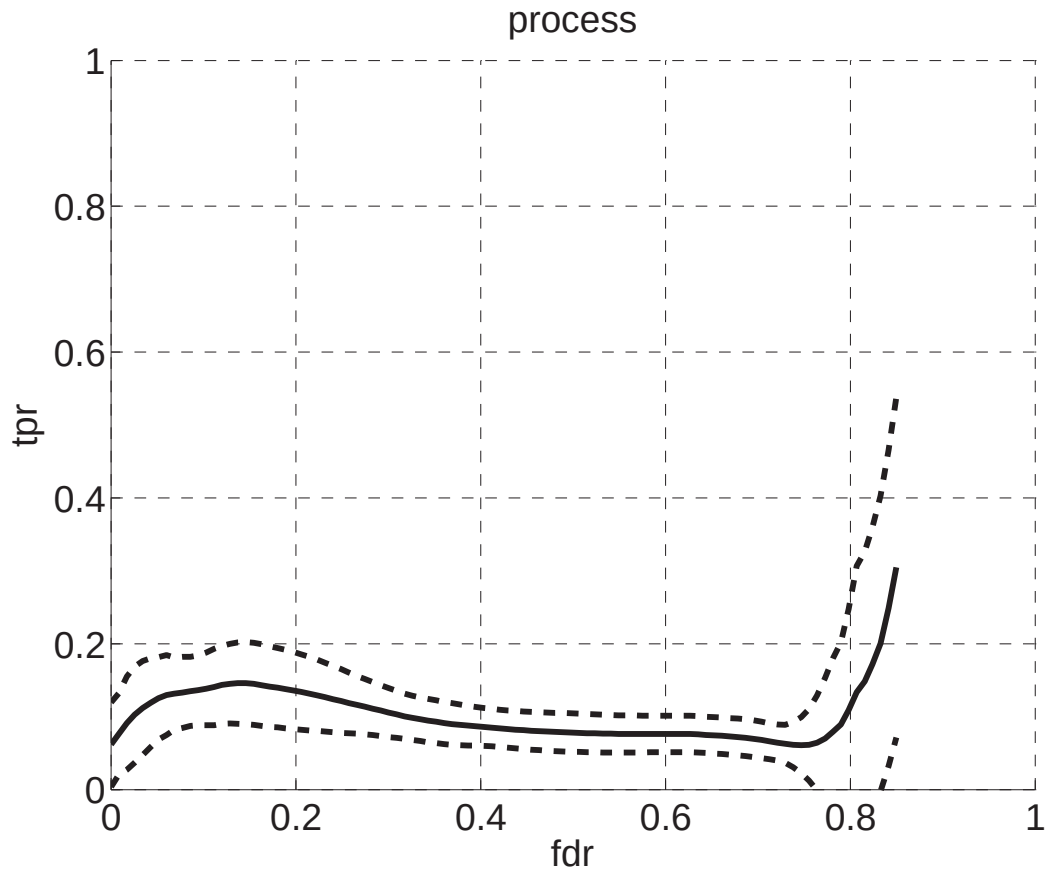

Figure 29: Operating characteristic: PROcess

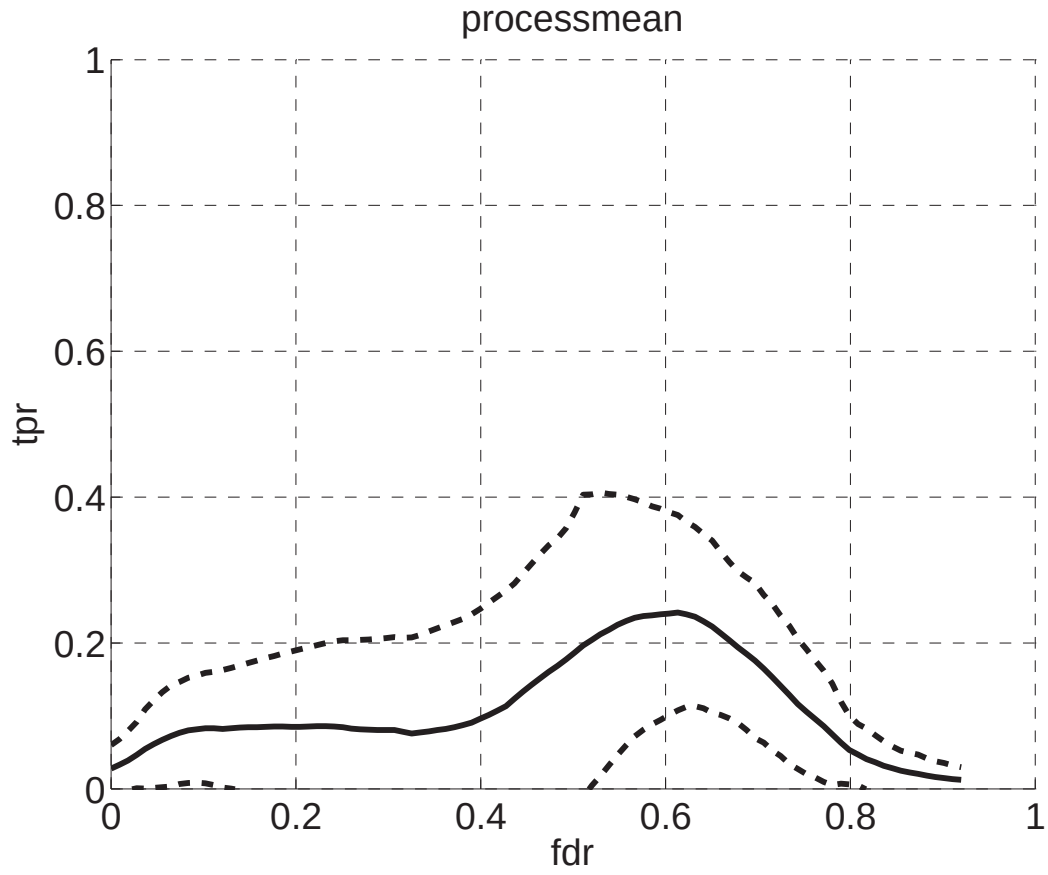

Figure 30: Operating characteristic: PROcess/mean spectrum

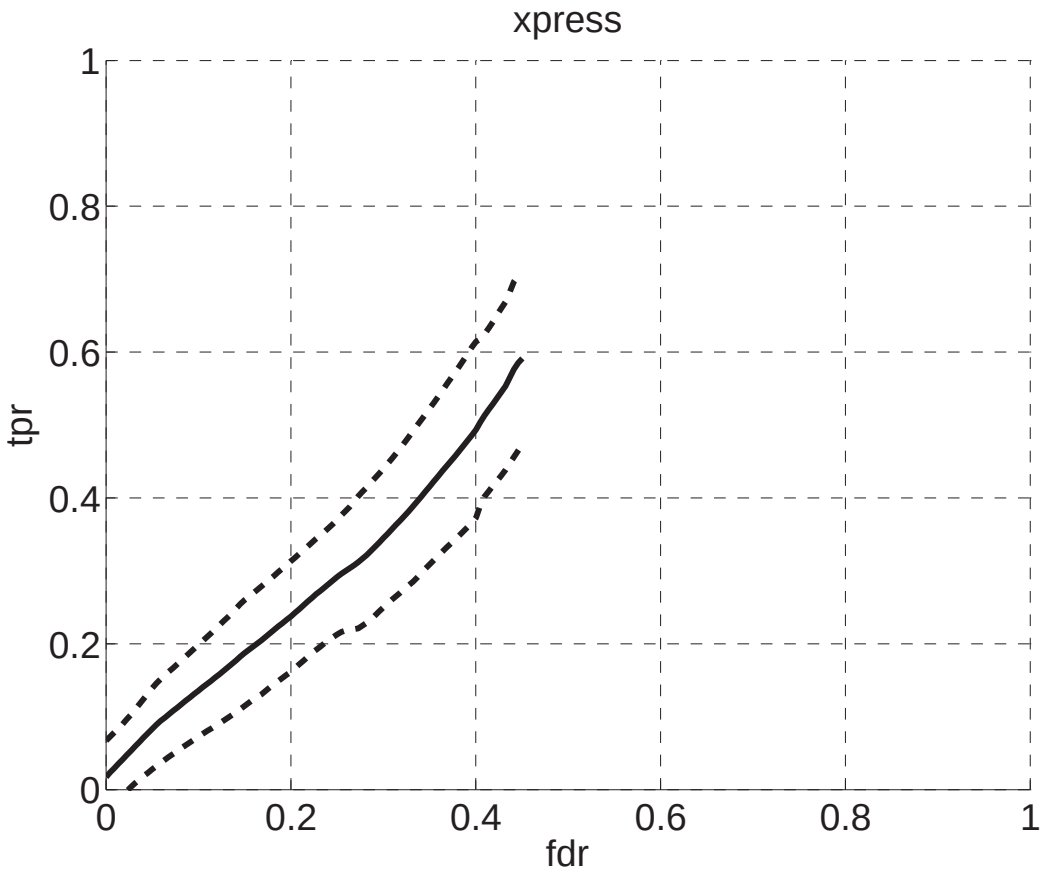

Figure 31: Operating characteristic: Ciphergen Express

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3: QUADRATIC VARIANCE MODELS FOR ADAPTIVE PREPROCESSING OF SELDI MASS SPECTROMETRY DATA

B.1 Reproducible computational research

LibSELDI and the associated MATLAB scripts and data necessary to reproduce the figures and tables shown in the main text are available for download from the following ftp server:

ftp server: `ftp://ftp.vincentemanuele.com/`
login and password: Available via email

The files are available in zipped format (Windows users) or tarballs (Unix/Linux users). Once the files are downloaded, unzip all of the files in the same directory and read the the provided README.txt and LICENSE.txt files. The software is provided under version 3 of the GNU Public License.

B.2 Relevant SELDI PBS IIc settings

We summarize the machine settings used to generate the buffer+matrix only QA/QC spectra. We have selected what we believe to be the most pertinent factors affecting the results seen in Fig 8 and 9. The lists in B.2.2 and B.2.3 are a summary of corresponding entries in the Ciphergen XML files produced. Additional parameters may be read directly out of the files. The QA/QC procedure used to produce the final BUFFER1 and BUFFER2 datasets is described in Section B.2.1.

B.2.1 QA/QC (outlier removal)

We examine all spectra generated for BUFFER1 using the quantile spectrum approach described in the main text. For a fixed t, outlier points are detected as any points falling

outside the the interval $[q_{25} - 1.5 \cdot IQR, q_{75} + 1.5 \cdot IQR]$, where q_{25}, q_{75}, and $IQR = q_{75} - q_{25}$ are the 25% quantile, 75% quantile, and inter-quartile range, respectively. This is done for all t. Any spectra containing one or more outlier points is declared an outlier and removed. This QA/QC procedure yields 183 high quality spectra for BUFFER1 and 114 high quality spectra for BUFFER2.

B.2.2 BUFFER1 settings (202 spectra, pre QA/QC)

- **spotProtocolInstructions:** Set high mass to 50000 Daltons, optimized from 3000 Daltons to 30000 Daltons. Set starting laser intensity to 185. Set starting detector sensitivity to 8. Focus by optimization center. Set Mass Deflector to 2000 Daltons. Set data acquisition method to Seldi Quantitation Set Seldi acquisition parameters 23. delta to 4. transients per to 12 ending position to 83. Set warming positions with 2 shots at intensity 195 and Do not include warming shots. Process sample.
- **ionFocusDelay:** 9.83e-007 s
- **deflectorMass:** 2000 Da
- **highMassCollected:** 50,000 Da
- **laserIntensityLow:** 185 (arbitrary units)
- **shotsFired:** 224
- **shotsKept:** 192
- **spotCorrectionEnabled:**
  - false (191 spectra)
  - true (11 spectra)

B.2.3 BUFFER2 settings (148 spectra, pre QA/QC)

- **spotProtocolInstructions:**

- Set high mass to 30000 Daltons, optimized from 3000 Daltons to 30000 Daltons. Set starting laser intensity to (multiple cases, see below). Set starting detector sensitivity to 7. Focus by optimization center. Set Mass Deflector to 2500 Daltons. Set data acquisition method to Seldi Quantitation Set Seldi acquisition parameters 20. delta to 4. transients per to 12 ending position to 80. Set warming positions with 2 shots at intensity 220 and Do not include warming shots. Process sample.
- **ionFocusDelay:** 9.83e-007 s
- **deflectorMass:** 2500 Da
- **highMassCollected:** 30,000 Da
- **laserIntensityLow:** (arbitrary units)
  - 185 (46 spectra), 190 (12 spectra), 195 (40 spectra), 200 (26 spectra), 210 (10 spectra), 215 (14 spectra)
- **shotsFired:** 224
- **shotsKept:** 192
- **spotCorrectionEnabled:**
  - true (148 spectra)
  - false (none)

B.3 MassSpecWavelet Code

Below we provide a code snippet to illustrate how MassSpecWavelet was used to calculate peak/protein predictions for each dataset over a wide range of snr settings in an efficient way. Note that MassSpecWavelet is implemented in the *R* computing language.

```
### ...,,, Calculate mean spectrum for directory
  print("Mean Spectrum Estimation\n")
  meanInt <- rowMeans(rawM$xtr)
```

```
  mzs <- rawM$mz

### ...,,, CWT
  print("CWT Calculation\n")
  wCoefs <- cwt( meanInt, scales = scales, wavelet = "mexh" )

  print("Analyzing CWT Result\n")
  wCoefs <- cbind( as.vector(meanInt), wCoefs)
  colnames(wCoefs) <- c(0,scales)
  localMax <- getLocalMaximumCWT(wCoefs)

### ...,,, Peak detection steps
  print("Looking for peaks/ridges\n")
  ridgeList <- getRidge(localMax)

  majorPeakInfo <- identifyMajorPeaks( meanInt, ridgeList, wCoefs,
                                       SNR.Th = 1, nearbyPeak=TRUE )

  # ...,,, Grab list of potential peaks and their SNRs
  potentialPeaks <- majorPeakInfo$potentialPeakIndex
  peakSNR <- majorPeakInfo$peakSNR[ names(potentialPeaks) ]
  uniqueSNR <- unique( peakSNR )
  sortedSNR <-sort( uniqueSNR, decreasing=TRUE )
  print("Done\n")
  # ... Create data structures for bookeeping
  peaks <- vector("list", length=length(sortedSNR) )
  attributes(peaks) <- list(names=rep("",length(sortedSNR)))
  parameters <- sortedSNR
```

```
# ...,,, Iterate over all SNRs and produce predictions.
for ( jsnr in 1:length(sortedSNR) ) {

  ## ... Get peaks with SNR at least this good
  thisSNR <- sortedSNR[ jsnr ]
  currentIndex <- peakSNR >= thisSNR
  currentPeaks <- potentialPeaks[ currentIndex ]

  peaks[[jsnr]] <- mzs[ currentPeaks ]

}
```

REFERENCES

[1] Adam, B.-L., Qu, Y., Davis, J. W., Ward, M. D., Clements, M. A., Cazares, L. H., Semmes, O. J., Schellhammer, P. F., Yasui, Y., Feng, Z., and Wright, G. L., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men.," *Cancer Res*, vol. 62, pp. 3609–3614, Jul 2002.

[2] Andrade, L. and Manolakos, E. S., "Signal background estimation and baseline correction algorithms for accurate DNA sequencing," *Journal of VLSI Signal Processing*, vol. 35, pp. 229–243, 2003.

[3] Antoniadis, A. and Sapatinas, T., "Wavelet shrinkage for natural exponential families with quadratic variance functions," *Biometrika*, vol. 88, no. 3, pp. 805–820, 2001.

[4] Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L.-C., and Coombes, K. R., "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, pp. 1667–1672, Sep 2003.

[5] Baggerly, K. A., Morris, J. S., and Coombes, K. R., "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments.," *Bioinformatics*, vol. 20, pp. 777–785, Mar 2004.

[6] Bantscheff, M., Duemelfeld, B., and Kuster, B., "An improved two-step calibration method for matrix-assisted laser desorption/ionization time-of-flight mass spectra for proteomics," *Rapid Commun Mass Spectrom*, vol. 16, no. 19, pp. 1892–1895, 2002.

[7] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B-Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

[8] Beyer, S., Walter, Y., Hellmann, J., Kramer, P.-J., Kopp-Schneider, A., Kroeger, M., and Ittrich, C., "Comparison of software tools to improve the detection of carcinogen induced changes in the rat liver proteome by analyzing SELDI-TOF-MS spectra.," *J Proteome Res*, vol. 5, pp. 254–261, Feb 2006.

[9] Bhanot, G., Alexe, G., Venkataraghavan, B., and Levine, A. J., "A robust meta-classification strategy for cancer detection from ms data.," *Proteomics*, vol. 6, pp. 592–604, Jan 2006.

[10] Bock, M. D., de Seny, D., Meuwis, M.-A., Chapelle, J.-P., Louis, E., Malaise, M., Merville, M.-P., and Fillet, M., "Challenges for biomarker discovery in body fluids using seldi-tof-ms.," *J Biomed Biotechnol*, vol. 2010, p. 906082, 2010.

[11] Borgia, J. A., Frankenberger, C., Kaiser, K., McCormack, S. E., Usha, L., Rotmensch, J., and Coon, J. S., "Serum biomarker discovery for ovarian serous carcinoma using novel proteomic methods," *J Clin Oncol (Meeting Abstracts)*, vol. 25, no. 18S, p. 16058, 2007.

[12] Carlson, S. M., Najmi, A., Whitin, J. C., and Cohen, H. J., "Improving feature detection and analysis of surface-enhanced laser desorption/ionization-time of flight mass spectra.," *Proteomics*, vol. 5, pp. 2778–2788, Jul 2005.

[13] Check, E., "Proteomics and cancer: running before we can walk?," *Nature*, vol. 429, pp. 496–497, Jun 2004.

[14] Chen, Y.-D., Zheng, S., Yu, J., and Hu, X., "Artificial neural networks analysis of surface-enhanced laser desorption/ionization mass spectra of serum protein pattern distinguishes colorectal cancer from healthy population.," *Clin Cancer Res*, vol. 10, pp. 8380–8385, Dec 2004.

[15] Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M., and Halfon, M. S., "Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset.," *Genome Biol*, vol. 6, no. 2, p. R16, 2005.

[16] Christian, N. P., Arnold, R. J., and Reilly, J. P., "Improved calibration of time-of-flight mass spectra by simplex optimization of electrostatic ion calculations," *Anal Chem*, vol. 72, pp. 3327–3337, Jul 2000.

[17] Chung, L., Clifford, D., Buckley, M., and Baxter, R. C., "Novel Biomarkers of Human Growth Hormone Action from Serum Proteomic Profiling Using Protein Chip Mass Spectrometry," *J Clin Endocrinol Metab*, vol. 91, no. 2, pp. 671–677, 2006.

[18] Coombes, K. R., Koomen, J. M., Baggerly, K. A., Morris, J. S., and Kobayashi, R., "Understanding the characteristics of mass spectrometry data through the use of simulation," *Cancer Informatics*, vol. 1, no. 1, pp. 41–52, 2005.

[19] Coombes, K. R., "Private communication." Discussion of Isotope Model and Calibration Error, May 2007.

[20] Coombes, K. R., Fritsche, H. A., Clarke, C., Chen, J.-N., Baggerly, K. A., Morris, J. S., Xiao, L.-C., Hung, M.-C., and Kuerer, H. M., "Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization.," *Clin Chem*, vol. 49, pp. 1615–1623, Oct 2003.

[21] Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C., and Kuerer, H. M., "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform.," *Proteomics*, vol. 5, pp. 4107–4117, Nov 2005.

[22] Corthals, G. L., Wasinger, V. C., Hochstrasser, D. F., and Sanchez, J. C., "The dynamic range of protein expression: a challenge for proteomic research.," *Electrophoresis*, vol. 21, pp. 1104–1115, Apr 2000.

[23] Cotter, R. J., *Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research*. ACS professional reference book, American Chemical Society, 1997.

[24] Cruz-Marcelo, A., Guerra, R., Vannucci, M., Li, Y., Lau, C. C., and Man, T.-K., "Comparison of algorithms for pre-processing of seldi-tof mass spectrometry data.," *Bioinformatics*, vol. 24, pp. 2129–2136, Oct 2008.

[25] Daubechies, I., *Ten lectures on wavelets*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1992.

[26] de Seny, D., Fillet, M., Meuwis, M.-A., Geurts, P., Lutteri, L., Ribbens, C., Bours, V., Wehenkel, L., Piette, J., Malaise, M., and Merville, M.-P., "Discovery of new rheumatoid arthritis biomarkers using the surface-enhanced laser desorption/ionization time-of-flight mass spectrometry ProteinChip approach.," *Arthritis Rheum*, vol. 52, pp. 3801–3812, Dec 2005.

[27] Diamandis, E. P., "Point: Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics?," *Clin Chem*, vol. 49, pp. 1272–1275, Aug 2003.

[28] Diamandis, E. P., "Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems.," *J Natl Cancer Inst*, vol. 96, pp. 353–356, Mar 2004.

[29] Diamandis, E. P., "Serum proteomic profiling by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry for cancer diagnosis: next steps.," *Cancer Res*, vol. 66, pp. 5540–5541, Jun 2006.

[30] Diamandis, E. P. and van der Merwe, D.-E., "Plasma protein profiling by mass spectrometry for cancer diagnosis: opportunities and limitations.," *Clin Cancer Res*, vol. 11, pp. 963–965, Feb 2005.

[31] Dietz, L. A., "Basic properties of electron multiplier ion detection and pulse counting methods in mass spectrometry," *Rev Sci Instrum*, vol. 36, pp. 1763–1770, December 1965.

[32] Domon, B. and Aebersold, R., "Mass spectrometry and protein analysis," *Science*, vol. 312, pp. 212–217, April 2006.

[33] Donoho, D. L. and Johnstone, I. M., "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[34] Donoho, D. L., "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data," in *In Proceedings of Symposia in Applied Mathematics*, pp. 173–205, American Mathematical Society, 1993.

[35] Du, P., Kibbe, W. A., and Lin, S. M., "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[36] Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P., "Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments," *Stat Sinica*, vol. 12, no. 1, pp. 111–139, 2002.

[37] Ebert, M. P. A., Meuer, J., Wiemer, J. C., Schulz, H.-U., Reymond, M. A., Traugott, U., Malfertheiner, P., and Rcken, C., "Identification of gastric cancer patients by serum protein profiling.," *J Proteome Res*, vol. 3, no. 6, pp. 1261–1266, 2004.

[38] Ekblad, L., Baldetorp, B., Fern, M., Olsson, H., and Bratt, C., "In-source decay causes artifacts in seldi-tof ms spectra.," *J Proteome Res*, vol. 6, pp. 1609–1614, Apr 2007.

[39] Emanuele, V. A. and Gurbaxani, B. M., "Benchmarking currently available seldi-tof ms preprocessing techniques.," *Proteomics*, vol. 9, pp. 1754–1762, Apr 2009.

[40] Emanuele, V. A. and Gurbaxani, B. M., "Quadratic variance models for adaptively preprocessing seld-tof mass spectrometry data," *BMC Bioinformatics*, vol. under review, 2010.

[41] Engwegen, J. Y. M. N., Helgason, H. H., Cats, A., Harris, N., Bonfrer, J. M. G., Schellens, J. H. M., and Beijnen, J. H., "Identification of serum proteins discriminating colorectal cancer patients and healthy controls using surface-enhanced laser desorption ionisation-time of flight mass spectrometry.," *World J Gastroenterol*, vol. 12, pp. 1536–1544, Mar 2006.

[42] Etzioni, R. and others, "The case for early detection.," *Nat. Rev. Cancer*, vol. 3, pp. 243–252, Apr 2003.

[43] Fawcett, T., "Roc graphs: Notes and practical considerations for data mining researchers," Tech Report HPL-2003-4, HP Laboratories, Palo Alto, CA, USA, 2003.

[44] Friesen, G. and others, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE T. Bio-Med Eng.*, vol. 37, pp. 85–98, January 1990.

[45] Fritsch, F. N. and Carlson, R. E., "Monotone piecewise cubic interpolation," *SIAM j. Numerical Analysis*, vol. 17, pp. 238–246, 1980.

[46] Fung, E. T. and Enderwick, C., "Proteinchip clinical proteomics: computational challenges and solutions," *BioTechniques*, vol. Suppl, pp. 34–38,40–41, March 2002.

[47] Fung, E. T., Yip, T.-T., Lomas, L., Wang, Z., Yip, C., Meng, X.-Y., Lin, S., Zhang, F., Zhang, Z., Chan, D. W., and Weinberger, S. R., "Classification of cancer types by measuring variants of host response proteins using seldi serum assays.," *Int J Cancer*, vol. 115, pp. 783–789, Jul 2005.

[48] Gatlin-Bunai, C. L., Cazares, L. H., Cooke, W. E., Semmes, O. J., and Malyarenko, D. I., "Optimization of maldi-tof ms detection for enhanced sensitivity of affinity-captured proteins spanning a 100 kda mass range.," *J Proteome Res*, vol. 6, pp. 4517–4524, Nov 2007.

[49] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biol*, vol. 5, no. 10, p. R80, 2004.

[50] Gloeckler Ries, L. A. and others, "Cancer Survival and Incidence from the Surveillance, Epidemiology, and End Results (SEER) Program," *Oncologist*, vol. 8, no. 6, pp. 541–552, 2003.

[51] Gobom, J., Mueller, M., Egelhofer, V., Theiss, D., Lehrach, H., and Nordhoff, E., "A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS," *Anal Chem*, vol. 74, pp. 3915–3923, Aug 2002.

[52] Gross, J. H., *Mass Spectrometry: A Textbook*. Berlin: Springer, 2004.

[53] Hack, C. A. and Benner, W. H., "A simple algorithm improves mass accuracy to 50-100 ppm for delayed extraction linear matrix-assisted laser desorption/ionization time-of-flight mass spectrometry," *Rapid Commun Mass Spectrom*, vol. 16, no. 13, pp. 1304–1312, 2002.

[54] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.

[55] Hernandez, P. and others, "Automated protein identification by tandem mass spectrometry: issues and strategies," *Mass Spectrom. Rev.*, vol. 25, no. 2, pp. 235–254, 2006.

[56] Ho, D. W. Y., Yang, Z. F., Wong, B. Y.-H., Kwong, D. L.-W., Sham, J. S.-T., Wei, W. I., and Yuen, A. P. W., "Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry serum protein profiling to identify nasopharyngeal carcinoma.," *Cancer*, vol. 107, pp. 99–107, Jul 2006.

[57] Hu, P., Le, W., Lim, S., Xing, B., Greenwood, C. M. T., and Beyene, J., "Serum Diagnosis of Chronic Fatigue Syndrome Using Array-based Proteomics," in *Oral Presenters Abstracts for Critical Assessment of Microarray Data Analysis (CAMDA 06)*, (Durham, NC, USA), pp. 50–59, 2006.

[58] Hutchens, T. W. and Yip, T., "New desorption strategies for the mass spectrometric analysis of macromolecules," *Rapid Commun Mass Spectrom*, vol. 7, no. 7, pp. 576–580, 1993.

[59] Hüttenhain, R., Malmström, J., Picotti, P., and Aebersold, R., "Perspectives of targeted mass spectrometry for protein biomarker verification.," *Curr Opin Chem Biol*, vol. 13, pp. 518–525, Dec 2009.

[60] Issaq, H. J., Conrads, T. P., Prieto, D. A., Tirumalai, R., and Veenstra, T. D., "Seldi-tof ms for diagnostic proteomics.," *Anal Chem*, vol. 75, pp. 148A–155A, Apr 2003.

[61] Issaq, H. J., Veenstra, T. D., Conrads, T. P., and Felschow, D., "The seldi-tof ms approach to proteomics: protein profiling and biomarker identification.," *Biochem Biophys Res Commun*, vol. 292, pp. 587–592, Apr 2002.

[62] Jeffries, N., "Algorithms for alignment of mass spectrometry proteomic data," *Bioinformatics*, vol. 21, pp. 3066–3073, Jul 2005.

[63] Jordanov, V. and Hall, D., "Digital peak detector with noise threshold," in *2002 IEEE Nuclear Science Symposium Conference Record*, vol. 1, pp. 140–142, November 2002.

[64] Juhasz, P., Vestal, M. L., and Martin, S. A., "On the initial velocity of ions generated by matrix-assisted laser desorption ionization and its effect on the calibration of delayed extraction time-of-flight mass spectra," *J Am Soc Mass Spectrom*, vol. 8, pp. 209–217, 1997.

[65] Kang, X., Xu, Y., Wu, X., Liang, Y., Wang, C., Guo, J., Wang, Y., Chen, M., Wu, D., Wang, Y., Bi, S., Qiu, Y., Lu, P., Cheng, J., Xiao, B., Hu, L., Gao, X., Liu, J., Wang, Y., Song, Y., Zhang, L., Suo, F., Chen, T., Huang, Z., Zhao, Y., Lu, H., Pan, C., and Tang, H., "Proteomic fingerprints for potential application to early diagnosis of severe acute respiratory syndrome.," *Clin Chem*, vol. 51, pp. 56–64, Jan 2005.

[66] Karas, M. and Hillenkamp, F., "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.," *Anal Chem*, vol. 60, pp. 2299–2301, Oct 1988.

[67] Kuerer, H. M., Coombes, K. R., Chen, J.-N., Xiao, L., Clarke, C., Fritsche, H., Krishnamurthy, S., Marcy, S., Hung, M.-C., and Hunt, K. K., "Association between ductal fluid proteomic expression profiles and the presence of lymph node metastases in women with breast cancer.," *Surgery*, vol. 136, pp. 1061–1069, Nov 2004.

[68] Lancashire, L., Schmid, O., Shah, H., and Ball, G., "Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis.," *Bioinformatics*, vol. 21, pp. 2191–2199, May 2005.

[69] Larman, M., Katz-Jaffe, M., Sheehan, C., and Gardner, D., "1,2-propanediol and the type of cryopreservation procedure adversely affect mouse oocyte physiology," *Hum. Reprod.*, vol. 22, no. 1, pp. 250–259, 2007.

[70] Lin, Y.-W., Lin, C.-Y., Lai, H.-C., Chiou, J.-Y., Chang, C.-C., Yu, M.-H., and Chu, T.-Y., "Plasma proteomic pattern as biomarkers for ovarian cancer.," *Int J Gynecol Cancer*, vol. 16 Suppl 1, pp. 139–146, 2006.

[71] Lundquist, M., Caspersen, M. B., Wikstrm, P., and Forsman, M., "Discrimination of francisella tularensis subspecies using surface enhanced laser desorption ionization mass spectrometry and multivariate data analysis.," *FEMS Microbiol Lett*, vol. 243, pp. 303–310, Feb 2005.

[72] Malyarenko, D. I., Cooke, W. E., Adam, B.-L., Malik, G., Chen, H., Tracy, E. R., Trosset, M. W., Sasinowski, M., Semmes, O. J., and Manos, D. M., "Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques.," *Clin Chem*, vol. 51, pp. 65–74, Jan 2005.

[73] Malyarenko, D. I., Cooke, W. E., Tracy, E. R., Drake, R. R., Shin, S., Semmes, O. J., Sasinowski, M., and Manos, D. M., "Resampling and deconvolution of linear time-of-flight records for enhanced protein profiling.," *Rapid Commun Mass Spectrom*, vol. 20, no. 11, pp. 1670–1678, 2006.

[74] Malyarenko, D., Cooke, W., Tracy, E., Trosset, M., Semmes, O., Sasinowski, M., and Manos, D., "Deconvolution filters to enhance resolution of dense time-of-flight survey spectra in the time-lag optimization range," *Rapid Commun Mass Spectrom*, vol. 20, no. 11, pp. 1661 – 9, 2006. deconvolution filters;dense time-of-flight survey spectral resolution;time-lag optimization range;time-domain filters;time-of-flight mass spectrometry signals;signal-to-noise ratio;filtering;smoothing;time series method;matrix-assisted laser desorption/ionization;biomolecules;nonlinear filters;.

[75] Mani, D. R. and Gillette, M., *New Generation of Data Mining Applications*, ch. Proteomic Data Analysis: Pattern Recognition for Medical Diagnosis and Biomarker Discovery. IEEE Press, 2005.

[76] Mazzulli, T., Low, D. E., and Poutanen, S. M., "Proteomics and severe acute respiratory syndrome (SARS): emerging technology meets emerging pathogen," *Clin Chem*, vol. 52, pp. 421–429, 2005.

[77] McLerran, D., Grizzle, W. E., Feng, Z., Bigbee, W. L., Banez, L. L., Cazares, L. H., Chan, D. W., Diaz, J., Izbicka, E., Kagan, J., Malehorn, D. E., Malik, G., Oelschlager, D., Partin, A., Randolph, T., Rosenzweig, N., Srivastava, S., Srivastava, S., Thompson, I. M., Thornquist, M., Troyer, D., Yasui, Y., Zhang, Z., Zhu, L., and Semmes, O. J., "Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias.," *Clin Chem*, vol. 54, pp. 44–52, Jan 2008.

[78] McLerran, D., Grizzle, W. E., Feng, Z., Thompson, I. M., Bigbee, W. L., Cazares, L. H., Chan, D. W., Dahlgren, J., Diaz, J., Kagan, J., Lin, D. W., Malik, G., Oelschlager, D., Partin, A., Randolph, T. W., Sokoll, L., Srivastava, S., Srivastava, S., Thornquist, M., Troyer, D., Wright, G. L., Zhang, Z., Zhu, L., and Semmes, O. J., "Seldi-tof ms whole serum proteomic profiling with imac surface does not reliably detect prostate cancer.," *Clin Chem*, vol. 54, pp. 53–60, Jan 2008.

[79] Meuleman, W., Engwegen, J. Y. M. N., Gast, M.-C. W., Wessels, L. F. A., and Reinders, M. J. T., "Analysis of mass spectrometry data using sub-spectra.," *BMC Bioinformatics*, vol. 10 Suppl 1, p. S51, 2009.

[80] Meuleman, W., Engwegen, J. Y., Gast, M.-C. W., Beijnen, J. H., Reinders, M. J., and Wessels, L. F., "Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (seldi) time-of-flight (tof) mass spectrometry data.," *BMC Bioinformatics*, vol. 9, p. 88, 2008.

[81] Morris, C. N., "Natural exponential families with quadratic variance functions," *The Annals of Statistics*, vol. 10, no. 1, pp. 65–80, 1982.

[82] Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R., "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum.," *Bioinformatics*, vol. 21, pp. 1764–1775, May 2005.

[83] Novikova, S. I., He, F., Cutrufello, N. J., and Lidow, M. S., "Identification of protein biomarkers for schizophrenia and bipolar disorder in the postmortem prefrontal cortex using SELDI-TOF-MS ProteinChip profiling combined with MALDI-TOF-PSD-MS analysis.," *Neurobiol Dis*, vol. 23, pp. 61–76, Jul 2006.

[84] Pan, Y.-Z., Xiao, X.-Y., Zhao, D., Zhang, L., Ji, G.-Y., Li, Y., Yang, B.-X., He, D.-C., and Zhao, X.-J., "Application of surface-enhanced laser desorption/ionization time-of-flight-based serum proteomic array technique for the early diagnosis of prostate cancer.," *Asian J Androl*, vol. 8, pp. 45–51, Jan 2006.

[85] Pang, R. T. K., Poon, T. C. W., Chan, K. C. A., Lee, N. L. S., Chiu, R. W. K., Tong, Y.-K., Chim, S. S. C., Sung, J. J. Y., and Lo, Y. M. D., "Serum amyloid a is not useful in the diagnosis of severe acute respiratory syndrome.," *Clin Chem*, vol. 52, pp. 1202–1204, Jun 2006.

[86] Panicker, G., Lee, D. R., and Unger, E. R., "Optimization of seldi-tof protein profiling for analysis of cervical mucous," *Journal of Proteomics*, vol. 71, no. 6, pp. 637 – 646, 2009.

[87] Panicker, G., Ye, Y., Wang, D., and Unger, E. R., "Characterization of the human cervical mucous proteome.," *Clin Proteomics*, vol. 6, pp. 18–28, Jun 2010.

[88] Papoulis, A. and Pillai, S. U., *Probability, random variables, and stochastic processes*. Mcgraw-Hill, 4 ed., 2002.

[89] Pawlik, T. M., Fritsche, H., Coombes, K. R., Xiao, L., Krishnamurthy, S., Hunt, K. K., Pusztai, L., Chen, J.-N., Clarke, C. H., Arun, B., Hung, M.-C., and Kuerer, H. M., "Significant differences in nipple aspirate fluid protein expression between healthy women and those with breast cancer demonstrated by time-of-flight mass spectrometry.," *Breast Cancer Res Treat*, vol. 89, pp. 149–157, Jan 2005.

[90] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A., "Use of proteomic patterns in serum to identify ovarian cancer.," *Lancet*, vol. 359, pp. 572–577, Feb 2002.

[91] Poon, T. C. W., "Opportunities and limitations of seldi-tof-ms in biomedical research: practical advices.," *Expert Rev Proteomics*, vol. 4, pp. 51–65, Feb 2007.

[92] Proakis, J. G., *Digital Communications*. McGraw Hill, 4 ed., 2000.

[93] Pusztai, L., Gregory, B. W., Baggerly, K. A., Peng, B., Koomen, J., Kuerer, H. M., Esteva, F. J., Symmans, W. F., Wagner, P., Hortobagyi, G. N., Laronga, C., Semmes, O. J., Wright, G. L., Drake, R. R., and Vlahou, A., "Pharmacoproteomic analysis of prechemotherapy and postchemotherapy plasma samples from patients receiving neoadjuvant or adjuvant chemotherapy for breast carcinoma.," *Cancer*, vol. 100, pp. 1814–1822, May 2004.

[94] Qu, Y., Adam, B.-L., Yasui, Y., Ward, M. D., Cazares, L. H., Schellhammer, P. F., Feng, Z., Semmes, O. J., and Wright, G. L., "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients.," *Clin Chem*, vol. 48, pp. 1835–1843, Oct 2002.

[95] Ressom, H. W., Varghese, R. S., Abdel-Hamid, M., Eissa, S. A.-L., Saha, D., Goldman, L., Petricoin, E. F., Conrads, T. P., Veenstra, T. D., Loffredo, C. A., and Goldman, R., "Analysis of mass spectral serum profiles for biomarker selection," *Bioinformatics*, vol. 21, pp. 4039–4045, Nov 2005.

[96] Rollin, D., Whistler, T., and Vernon, S. D., "Laboratory methods to improve seldi peak detection and quantitation.," *Proteome Sci*, vol. 5, p. 9, 2007.

[97] Sauve, A. C. and Speed, T. P., "Normalization, baseline correction and alignment of high-throughput mass spectrometry data," in *Workshop on Genomic Signal Processing (GENSIPS)*, (Baltimore, MD), May 26-27 2004.

[98] Savitzky, A. and Golay, M. J. E., "Smoothing and differentiation of data by simplified least squares procedures," *Anal Chem*, vol. 36, pp. 1627–1639, July 1964.

[99] Schwab, M., Karrenbach, N., and Claerbout, J., "Making scientific computations reproducible," *Computing in Science & Engineering*, vol. 2, no. 6, pp. 61–67, 2000.

[100] Semmes, O. J., Feng, Z., Adam, B.-L., Banez, L. L., Bigbee, W. L., Campos, D., Cazares, L. H., Chan, D. W., Grizzle, W. E., Izbicka, E., Kagan, J., Malik, G., McLerran, D., Moul, J. W., Partin, A., Prasanna, P., Rosenzweig, J., Sokoll, L. J., Srivastava, S., Srivastava, S., Thompson, I., Welsh, M. J., White, N., Winget, M., Yasui, Y., Zhang, Z., and Zhu, L., "Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. assessment of platform reproducibility.," *Clin Chem*, vol. 51, pp. 102–112, Jan 2005.

[101] Siuzdak, G., *The Expanding Role of Mass Spectrometry in Biotechnology*. MCC Press, 2003.

[102] Sköld, M., Rydén, T., Samuelsson, V., Bratt, C., Ekblad, L., Olsson, H., and Baldetorp, B., "Regression analysis and modelling of data acquisition for seldi-tof mass spectrometry.," *Bioinformatics*, vol. 23, pp. 1401–1409, Jun 2007.

[103] Sorace, J. M. and Zhan, M., "A data review and re-assessment of ovarian cancer serum proteomic profiling.," *BMC Bioinformatics*, vol. 4, p. 24, Jun 2003.

[104] Steen, H. and Mann, M., "The abc's (and xyz's) of peptide sequencing," *Nat. Rev. Mol. Cell Bio.*, vol. 5, pp. 699–711, 2004.

[105] Stekel, D., *Microarray Bioinformatics*. Cambridge University Press, 2003.

[106] Tan, C. S., Ploner, A., Quandt, A., Lehti, J., and Pawitan, Y., "Finding regions of significance in SELDI measurements for identifying protein biomarkers.," *Bioinformatics*, vol. 22, pp. 1515–1523, Jun 2006.

[107] Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T., "Protein and Polymer Analyses up to m/z 100,000 by Laser Ionization Time-of-flight Mass Spectrometry," *Rapid Commun Mass Spectrom*, vol. 2, no. 8, pp. 151–153, 1988.

[108] Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T., "Sample classification from protein mass spectrometry, by 'peak probability contrasts'," *Bioinformatics*, vol. 20, pp. 3034–3044, Nov 2004.

[109] Timms, J. F., Arslan-Low, E., Gentry-Maharaj, A., Luo, Z., T'Jampens, D., Podust, V. N., Ford, J., Fung, E. T., Gammerman, A., Jacobs, I., and Menon, U., "Preanalytic influence of sample handling on seldi-tof serum protein profiles.," *Clin Chem*, vol. 53, pp. 645–656, Apr 2007.

[110] Tuszynski, J., *Processing & Classification of Protein Mass Spectra (SELDI) Data: The caMassClass Package*, April 2006.

[111] van der Ziel, A., *Noise in Measurements*. Wiley-Interscience, 1976.

[112] Viv-Truyols, G., Torres-Lapasi, J. R., van Nederkassel, A. M., Heyden, Y. V., and Massart, D. L., "Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part i: peak detection.," *J Chromatogr A*, vol. 1096, pp. 133–145, Nov 2005.

[113] Viv-Truyols, G., Torres-Lapasi, J. R., van Nederkassel, A. M., Heyden, Y. V., and Massart, D. L., "Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals part ii: peak model and deconvolution algorithms.," *J Chromatogr A*, vol. 1096, pp. 146–155, Nov 2005.

[114] Vlahou, A., Schorge, J. O., Gregory, B. W., and Coleman, R. L., "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data.," *J Biomed Biotechnol*, vol. 2003, no. 5, pp. 308–314, 2003.

[115] Wagner, M., Naik, D. N., Pothen, A., Kasukurti, S., Devineni, R. R., Adam, B.-L., Semmes, O. J., and Wright, G. L., "Computational protein biomarker prediction: a case study for prostate cancer.," *BMC Bioinformatics*, vol. 5, p. 26, Mar 2004.

[116] Watson, J. T., *Introduction to Mass Spectrometry*. New York: Raven Press, 1985.

[117] Wegdam, W., Moerland, P. D., Buist, M. R., van Themaat, E. V. L., Bleijlevens, B., Hoefsloot, H. C. J., de Koster, C. G., and Aerts, J. M. F. G., "Classification-based comparison of pre-processing methods for interpretation of mass spectrometry generated clinical datasets.," *Proteome Sci*, vol. 7, p. 19, 2009.

[118] Wei, W., Martin, A., Johnson, P. J., and Ward, D. G., "10 years of seldi: What have we learnt?," *Current Proteomics*, vol. 7, pp. 15–25(11), April 2010.

[119] White, C. N., Chan, D. W., and Zhang, Z., "Bioinformatics strategies for proteomic profiling.," *Clin Biochem*, vol. 37, pp. 636–641, Jul 2004.

[120] White, C. N., Zhang, Z., and Chan, D. W., "Quality control for SELDI analysis.," *Clin Chem Lab Med*, vol. 43, no. 2, pp. 125–126, 2005.

[121] Wiesner, A., "Detection of tumor markers with proteinchip® technology," *Curr Pharm Biotechnol*, vol. 5, pp. 45–67, 2004.

[122] Wilkins, M. R. and others, "Guidelines for the next 10 years of proteomics," *Proteomics*, vol. 6, no. 1, pp. 4–8, 2006.

[123] Wolski, W. E., Lalowski, M., Jungblut, P., and Reinert, K., "Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants.," *BMC Bioinformatics*, vol. 6, p. 203, 2005.

[124] Wong, J. W. H., Durante, C., and Cartwright, H. M., "Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets.," *Anal Chem*, vol. 77, pp. 5655–5661, Sep 2005.

[125] Wong, J. W. H., Cagney, G., and Cartwright, H. M., "Specalign–processing and alignment of mass spectra datasets.," *Bioinformatics*, vol. 21, pp. 2088–2090, May 2005.

[126] Woroniecki, R. P., Orlova, T. N., Mendelev, N., Shatat, I. F., Hailpern, S. M., Kaskel, F. J., Goligorsky, M. S., and O'Riordan, E., "Urinary Proteome of Steroid-Sensitive and Steroid-Resistant Idiopathic Nephrotic Syndrome of Childhood.," *Am J Nephrol*, vol. 26, pp. 258–267, Jun 2006.

[127] Wulfkuhle, J. D., Liotta, L. A., and Petricoin, E. F., "Proteomic applications for the early detection of cancer," *Nat Rev Cancer*, vol. 3, pp. 267–275, 2003.

[128] Xu, X.-Q., Leow, C. K., Lu, X., Zhang, X., Liu, J. S., Wong, W.-H., Asperger, A., Deininger, S., and Leung, H.-C. E., "Molecular classification of liver cirrhosis in a rat model by proteomics and bioinformatics.," *Proteomics*, vol. 4, pp. 3235–3245, Oct 2004.

[129] Yang, S.-Y., Xiao, X., Zhang, W.-G., Zhang, L.-J., Zhang, W., Chen, G., and He, D.-C., "Application of serum SELDI proteomic patterns in diagnosis of lung cancer.," *BMC Cancer*, vol. 5, p. 83, 2005.

[130] Yasui, Y., McLerran, D., Adam, B.-L., Winget, M., Thornquist, M., and Feng, Z., "An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers," *J Biomed Biotechnol*, vol. 2003, no. 4, pp. 242–248, 2003.

[131] Yasui, Y., Pepe, M., Thompson, M. L., Adam, B.-L., Wright, G. L., Qu, Y., Potter, J. D., Winget, M., Thornquist, M., and Feng, Z., "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection.," *Biostatistics*, vol. 4, pp. 449–463, Jul 2003.

[132] Yu, J. S., Ongarello, S., Fiedler, R., Chen, X. W., Toffolo, G., Cobelli, C., and Trajanoski, Z., "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data.," *Bioinformatics*, vol. 21, pp. 2200–2209, May 2005.

[133] Zhai, R., Su, S., Lu, X., Liao, R., Ge, X., He, M., Huang, Y., Mai, S., Lu, X., and Christiani, D., "Proteomic profiling in the sera of workers occupationally exposed to arsenic and lead: identification of potential biomarkers.," *Biometals*, vol. 18, pp. 603–613, Dec 2005.

[134] Zhang, X., Lu, X., Shi, Q., Xu, X.-Q., Leung, H.-C. E., Harris, L. N., Iglehart, J. D., Miron, A., Liu, J. S., and Wong, W. H., "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data.," *BMC Bioinformatics*, vol. 7, p. 197, 2006.

VITA

Vincent A. Emanuele II was born in Olney, Maryland on February 19, 1979. He received the B.E. degree in electrical engineering at the University of Delaware cum laude and with distinction in 2002. In 2004, he completed the M.S. in electrical and computer engineering from the Georgia Institute of Technology. Since then, he has been working towards the Ph.D. degree in electrical and computer engineering at the Georgia Institute of Technology. From 2006 - 2010, he was a part of the guest researcher program at the Centers for Disease Control and Prevention in Atlanta, GA through the ORISE fellowship program. Since May 2010, he has been working full time as an associate service fellow at the CDC. Vincent's research interests include proteomics, bioinformatics, machine learning, statistical signal processing, and biotechnology.